

**A Thesis Submitted for the Degree of PhD at the University of Warwick**

**Permanent WRAP URL:**

<http://wrap.warwick.ac.uk/109025>

**Copyright and reuse:**

This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it.

Our policy information is available from the repository home page.

For more information, please contact the WRAP Team at: [wrap@warwick.ac.uk](mailto:wrap@warwick.ac.uk)

# THE BRITISH LIBRARY

BRITISH THESIS SERVICE

TITLE	BAYESIAN INFERENCE ON NON-STATIONARY DATA
AUTHOR	Giovanni AMISANO
DEGREE	Ph.D
AWARDING BODY	Warwick University
DATE	1995
THESIS NUMBER	DX193331

THIS THESIS HAS BEEN MICROFILMED EXACTLY AS RECEIVED

The quality of this reproduction is dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction. Some pages may have indistinct print, especially if the original papers were poorly produced or if awarding body sent an inferior copy. If pages are missing, please contact the awarding body which granted the degree.

Previously copyrighted materials (journals articles, published texts etc.) are not filmed.

This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with its author and that no information derived from it may be published without the author's prior written consent.

Reproduction of this thesis, other than as permitted under the United Kingdom Copyright Designs and Patents Act 1988, or under specific agreement with the copyright holder, is prohibited.

C 1.

**BAYESIAN INFERENCE ON  
NON-STATIONARY DATA**

**Giovanni Amisano**

Department of Economics,  
University of Warwick

September 1995

Thesis submitted for the award of the Ph.D. degree in  
economics

**To my parents with love and gratitude**

*Two roads diverged in the yellow wood,  
and I,*

*I took the road less traveled by*

*And that has made all the difference*

## **Table of Contents:**

<b>Acknowledgements</b>	<b>v</b>
<b>Summary</b>	<b>vi</b>
<b>Chapter 1: Introduction and Outline</b>	<b>1</b>
<b>PART I: The Univariate Analysis of Non-Stationary Time Series</b>	<b>7</b>
<b>Chapter 2: The Frequentist Approach to Non-stationarity in univariate Time Series Analysis</b>	<b>7</b>
[2.0] An Overview of the Chapter	7
[2.1] Integrated versus trend stationary processes	7
[2.2] Inference on trend stationary and integrated univariate processes	13
[2.3] Some General Considerations About Unit Root Testing in a Classical Framework	25
<b>Chapter 3: The Bayesian Approach to Time Series Analysis: Problems and Methods.</b>	<b>28</b>
[3.0] An Overview of the Chapter	28
[3.1] General Philosophy of the Bayesian Approach	28
[3.2] Bayesian Inferential Techniques	33
[3.3] The specification of the priors	40
[3.4] Computational Problems and Technical Solutions	46
[3.4.a] Approximations	47
[3.4.b] Numerical integration	48
[3.4.c] Monte Carlo Integration and Importance Sampling	49
[3.4.d] Markov Chain Monte Carlo: Gibbs Sampling and Metropolis- Hastings Algorithms	56
[3.4.e] Measuring the accuracy of MC Monte Carlo estimates	63
[3.5] Bayesian Analysis of Non Stationary Univariate Models	65

[3.6] Ignorance Priors in Time Series Models	74
<b>Chapter 4: Bayesian Analysis of Integration at Different Frequencies in Quarterly Data</b>	<b>78</b>
[4.0] An Overview of the Chapter	78
[4.1] General Features of the Model	80
[4.2] The Specification of the Priors	84
[4.3] The Joint Posterior Distribution	87
[4.4] The conditional posterior distributions	90
[4.5] A Convenient Description of the Posterior Odds Ratio	91
[4.6] An Application	97
[4.7] Conclusion	105
Appendix [4.A]: Proofs of distributional results	105
Appendix [4.B] : Rejection Sampling from the Conditional Posterior Distributions	110
Appendix [4.C] : Proofs of the Smooth Transition Results	113
<b>PART II: The Multivariate Analysis of Non Stationary Time Series</b>	<b>120</b>
<b>Chapter 5: Non Stationarity in Multivariate Time Series Analysis.</b>	<b>120</b>
<b>The Classical Approach</b>	
[5.0] An Overview of the Chapter	120
[5.1] Spurious Regression and Cointegration	121
[5.2] Representation and identification issues	123
[5.3] Estimation Issues	128
[5.4] Interpretation of the Cointegrating Coefficients	136
[5.5] Asymptotic Distributions of the Parameter Estimates	139
[5.6] Finite Sample Properties	141

<b>Chapter 6: Bayesian Inference in Cointegrated Systems</b>	<b>145</b>
[6.0] An Overview of the Chapter	145
[6.1] Motivations	146
[6.2] The Model	148
[6.3] The Prior Distribution	150
[6.4] The Joint Posterior Distribution	152
[6.5] Inference on Cointegration Rank	159
[6.6] Testing Restrictions on the Cointegration Space	162
[6.7] Some Applications	165
[6.7.1] A simulated data set example	166
[6.7.2] The Danish Money Demand Example	168
[6.7.3] The Finnish Money Demand Example	172
[6.7.4] The UK PPP/UIP Example	175
[6.8] Conclusion	179
<b>Chapter 7: Concluding remarks</b>	<b>191</b>
<b>References</b>	<b>196</b>

#### **List of tables and illustrations**

Table 4.1	100
Table 4.2	100
Table 4.3	101
Table 4.4	104
Figure 4.1	115
Figure 4.2	116
Figure 4.3	117
Figure 4.4	118
Figure 4.5	119
Table 6.1	167

Table 6.2.1	171
Table 6.2.2	172
Table 6.3.1	174
Table 6.3.2	174
Table 6.4.1	178
Table 6.4.2	179
Figure 6.2.1	180
Figure 6.2.2	180
Figure 6.2.3	181
Figure 6.2.4	181
Figure 6.2.5	182
Figure 6.3.1	182
Figure 6.3.2	183
Figure 6.3.3	183
Figure 6.3.4	184
Figure 6.3.5	184
Figure 6.3.6	185
Figure 6.4.1	185
Figure 6.4.2	186
Figure 6.4.3	186
Figure 6.4.4	187
Figure 6.4.5	187
Figure 6.4.6	188
Figure 6.4.7	188
Figure 6.4.8	189
Figure 6.4.9	189
Figure 6.4.10	190
Figure 6.4.11	190



## Aknowledgments

I would like to express my sincere gratitude to all the people I bothered with my questions. I greatly benefitted from discussions with Rocco Mosconi, Mario Seghelini, Jeremy Smith and Sanjay Yadav. I also received valuable suggestions from Frank Kleibergen, John Geweke and Hermann van Dijk at different stages of my work.

Carlo Giannini deserves all my gratitude for having been my first and great econometrics teacher. He started me off as an econometrician, encouraged me to go and study abroad, and gave me constant scientific inspiration throughout my career. *Grazie Carlo, sei un fratello.*

Kenneth Wallis, my supervisor, let me have the benefit of all his experience and competence, and gave me his warm encouragement in several difficult stages in the preparation of this thesis. I sincerely thank him for that, and I apologise for having so often tested his infinite patience.

The only non-econometrician person to appear in this list is my wife Anne who did not know what "Bayesian" and "unit root" meant before she met me. I am not sure she even knows now, but she does not seem to mind.

## Summary

*This thesis argues in favour of Bayesian techniques for the analysis of non-stationary linear time series. The main motivations are to avoid using asymptotic results and to explicitly incorporate prior beliefs, where they exist.*

*The properties of univariate and multivariate unit root models, and the available frequentist inferential results are described. Some problems in their applications are highlighted: the discrepancies between asymptotic and finite sample properties and the role of the deterministic components in determining the reference asymptotic distributions.*

*The advantages and disadvantages of Bayesian techniques are then examined with the recent developments in the Monte Carlo integration by Markov Chain sampling. Two case studies are conducted with the aim of providing evidence of the applicability of Bayesian techniques.*

*The first of these cases develops a procedure to test for seasonal and/or zero frequency unit roots in quarterly series. A new parameterisation is provided and the priors implemented are discussed and justified. The analysis relies on a Gibbs sampling scheme. The inferential technique used is the evaluation of posterior odds ratios. These ratios are defined as posterior expectations of functions of the parameters, and therefore can be consistently estimated. The procedure is applied to some UK variables. The results are robust with respect to different prior distributions, and conflict with some conclusions reached by using classical asymptotic unit root tests.*

*The second case study develops a Bayesian procedure to conduct inference in cointegrated systems. Inference regards the number of cointegrating relationships and their structural interpretation, and is based on the evaluation of highest posterior density confidence intervals.*

*The procedure is applied to three VAR systems: Danish and Finnish money demands, and UK exchange rate data. Interesting results emerge, showing significant differences with their frequentist counterparts. All these results are robust with respect to different priors.*

## **Chapter 1: Introduction and Outline**

Weakly stationarity processes, processes whose first and second order moments do not vary over time, have played a central role in the traditional econometric analysis of linear time series data. Unfortunately, assumptions of constancy of moments seem at odds with most observed economic time series.

Different alternative models of non-stationarity behaviour have been proposed, implying radically different long-run properties for the series under study. A simple way to account for non-stationarity is to assume that the process is stationary around a deterministic trend. Such a process has constant second order moments and shocks have only a transitory effect on it. Alternatively, it is very often assumed that one or more unit roots are present in the autoregressive representation. A unit root process is characterised by a growing variance, and by the fact that shocks have a permanent effect on the level of the series.

Economic theory has provided models implying the presence of unit roots in many macroeconomic aggregates. General equilibrium business cycle models emphasise the role of persistent shocks. Intertemporal quadratic utility maximisation leads to non-stationarity for individual consumption. The efficient market hypothesis implies an exploding variance for the asset price forecast error, as the forecasting horizon grows.

In the last two decades the statistical analysis of linear time series has made enormous progress by providing the researcher with the technical tools necessary to deal with unit root processes, and to discriminate between different models of non-stationarity. The asymptotic distributions of the parameters estimates for unit root autoregressive processes have been thoroughly investigated and some interesting features have been revealed.

First of all, these distributions are non-standard, since they are complicated functionals of Brownian motion processes. The exact form of these functionals depends on which deterministic components are included in the estimated model and in the true data generation mechanism. On the basis of these new asymptotic results, some testing procedures have been developed to ascertain the presence of unit roots in observed time series and therefore to discriminate between competing models of non-stationary behaviour.

As in the analysis of the long-run properties of univariate processes, also the phenomenon of seasonality can be explained on the basis of different competing models. A first possibility is to account for seasonality by introducing a set of seasonal dummy variables. A second possibility is that unit roots at seasonal frequencies are present in the autoregressive polynomial. This particular form of seasonality requires the application of an adequate filter to induce stationarity and raises the issue of the occurrence of common stochastic seasonality patterns in multivariate time series. Seasonal unit root tests have been developed, in order to discriminate between competing ways to model seasonality. As the zero-frequency unit root tests, these testing procedures are based on non-standard asymptotic distributional results.

In the analysis of multivariate time series, the problem of the interpretation of results of regressions among non-stationary variables is directly connected to the "spurious regression" problem. The notion of spurious regression relates to a regression among non stationary variables, when good measures of fit may be found even in the absence of any direct links among the variables.

In many cases, though, true long-run relationships do exist among non-stationary variables. Long-run relationships are particularly interesting because they relate to the notion of equilibrium links among sets of economic variables. The widely popular concept of cointegration directly refers to the existence of long-run relationships. Cointegration is defined as rank deficiency in the matrix of the long-

run multipliers in the autoregressive representation of a vector series. The rank of this matrix gives the number of stationary variables generated as independent linear combinations of the non-stationary series being considered. Each one of these stationary variables can be interpreted as deviations from a corresponding long-run relationship.

During the last few years, new inferential techniques have been proposed in order to analyse potentially cointegrated vector series. Inference mainly regards the number of cointegrating relationships and their structural interpretation as equilibrium relationships. The asymptotic distributional properties of estimators and testing procedures being used in this regard are again non-standard and depend on which deterministic components are thought to be present in the 'true' data generation process.

In synthesis, the analysis of univariate and multivariate inferential properties of unit root processes led to a great advancement in the statistical foundations of econometric modelling, allowing proper treatment of non-stationary data.

Unfortunately, these new inferential results present the applied researcher with some unpleasant features. Many Monte Carlo studies have revealed that the finite sample inferential properties of non-stationary models can be radically different to their known asymptotic counterparts. In most macroeconometric applications, where the typical sample size is well below 100 observations, reliance on inappropriate inferential results is extremely likely. In addition, the sensitivity of the scaled asymptotic distribution to the deterministic part of the model causes further complications, since the researcher cannot be sure about the correctness of the model specification in this regard.

Moreover, it is evident that the finite sample behaviour of observed series can be explained almost equally well by unit root and by stationary near-unit root processes. This problem of observational equivalence clearly generates very bad finite sample performances of the unit root and cointegration rank tests.

For these reasons, recent contributions in the analysis of non-stationary series have suggested that resorting to a Bayesian inferential framework could yield more sensible results.

Bayesian analysis presents three main advantages. First of all, uncertainty about the parameters can be directly measured on the basis of their posterior distributions, and no reference to asymptotic results is ever necessary. Secondly, a Bayesian approach requires a clear statement of the researcher's beliefs, in the form of the specification of a prior distribution for the parameters. This does not happen in applications of the frequentist inferential procedures, where prior beliefs, though unstated, are often incorporated. Thirdly, unlike the classical Neyman-Pearson apparatus, Bayesian model selection techniques are fully consistent, since the probabilities of picking a wrong model go to zero as more sample information becomes available.

On the other hand, Bayesian methods present two primary disadvantages. The first one is related to the necessity of providing a prior distribution. While most Bayesian researchers agree on the irrelevance of the issue of how to represent prior ignorance, a central problem in Bayesian analysis is how to render results universally acceptable, although they are clearly based on personal convictions.

The second disadvantage is a computational one. In fact, the typical results of Bayesian analysis can be defined as posterior expectations of certain functions of interest; these functions need to be integrated with respect to the posterior probability density function. Excluding only a narrow class of cases, this integration is almost always analytically unfeasible.

A satisfactory way of overcoming the first disadvantage is to assess the sensitivity of the results with respect to the choice of the prior distribution. This can be done by providing results corresponding to different alternative priors.

As for the computational difficulties, the Monte Carlo principle can be used to perform analytically unfeasible integration. Monte Carlo integration delivers

consistent estimates of the posterior expectations being studied, but in this context "consistency" refers to the number of steps in the simulations, which can be as high as desired, and not to the sample size which is seldom the result of the researcher's choice.

The main problem is then that of efficiently simulating the relevant posterior distributions, and this goal can be satisfactorily achieved in most econometric applications by using Markov chain sampling schemes.

This dissertation intends to provide examples of how Bayesian analysis can be efficiently used to conduct inference on non-stationary series. These examples take the form of new applications of posterior inference techniques to series which potentially have unit roots. The thesis deals with both univariate and multivariate issues and is structured as follows.

Chapter 2 explains the different properties of unit root and trend stationary univariate processes, reviews the frequentist inferential results available for univariate unit root process, and discusses the properties of the main testing procedures to discriminate between the two competing ways to account for non-stationarity.

In Chapter 3 the main characteristics of the Bayesian methodology are examined. Computational problems and technical solutions are discussed at length, with special emphasis on Markov chain Monte Carlo methods. The literature on the application of Bayesian inferential techniques on unit root processes is surveyed.

Chapter 4 describes a Bayesian inferential methodology to test for the presence of unit roots at seasonal frequencies. The technique is based on the evaluation of posterior odds ratios by means of a Gibbs sampling scheme. An application on a set of UK series is presented.

Chapter 5 discusses the problems of how to determine the number of cointegrating vectors and how to give them a structural interpretation. The testing procedures

based on maximum likelihood estimation are discussed together with their asymptotic and finite sample performances.

Chapter 6 describes a Bayesian procedure to test for the cointegrating rank based on a Gibbs sampling scheme. Having determined the rank, other Bayesian procedures are proposed to test for the over-identifying restrictions on the cointegration space. Some applications are presented on money demand and exchange rate examples.

Chapter 7 contains some conclusive considerations on the evidence gathered in this thesis and connects the present work with the on-going research in the area.



## **PART I: The Univariate Analysis of Non-Stationary Time Series**

### **Chapter 2: The Frequentist Approach to Non-stationarity in univariate Time Series Analysis.**

#### **[2.0] An Overview of the Chapter.**

This chapter contains an overview of the issue of non-stationarity in univariate models as seen from a frequentist point of view. In the first section the concepts of trend and difference stationary processes are introduced, and their different modelling properties are highlighted. In Section [2.2], I describe the existing inferential procedures designed to detect the presence of unit roots. Section [2.3] contains a brief summary of the main problems encountered in using the techniques surveyed in the previous section, and provides the main motivations supporting the adoption of a Bayesian inferential strategy.

#### **[2.1] Integrated versus trend stationary processes.**

Many macroeconomic time series show evident trending patterns. The theoretical justification for such non-stationary behaviour is often related to the occurrence of phenomena like technological progress, increases in population and in the capital stock, in short to all the forces supposed to drive the key economic variables over time in the long run. However, the tools of time series analysis are mainly based on the assumption of weak stationarity of the series under study, that is, the assumption that first and second order moments are time-invariant. Therefore, how to deal properly with the observed non-stationarity is an important question, and two distinct approaches have been developed in the literature. First, a reasonably

simple and straightforward procedure could be to assume the presence of a deterministic trend, and detrend the data accordingly by means of a simple regression; having performed such a transformation, one could then focus on the resulting series and model the remaining dynamics by means of the available techniques that rely on stationarity. Thus, it is implicitly assumed that the series is affected by a steady growth pattern, around which it fluctuates due to the transitory effects of disturbances. Such series is said to be 'trend stationary', i.e. to possess a stationary invertible ARMA representation once the trend has been removed. In the case of a linear trend, we have:

$$\rho(L)(y_t - \mu - \delta t) = \theta(L) e_t \quad (1)$$

where  $e_t$  is i.i.d. distributed with mean 0 and variance  $\sigma_e^2$ , and  $\rho(L)$  and  $\theta(L)$  are respectively stationary and invertible, thus in particular  $y_t - \mu - \delta t$  has bounded variance.

In this framework, the long run is accounted for in a completely deterministic way (i.e. it is predictable with zero error), and what remains is produced as the dynamic response to the realisation of a series of random disturbances. These shocks have an effect that vanishes in the long run. In fact, the zero-mean stationary process  $z_t = y_t - \mu - \delta t$  admits a Old representation:

$$z_t = [\theta(L)/\rho(L)] e_t = c(L) e_t, \quad (2)$$

$$\sum_{j=0}^{\infty} c_j^2 < \infty.$$

Given the square summability of MA coefficients for a stationary process, we see that the effect of a random shock on the levels of the series tends to vanish as time elapses:

$$\lim_{k \rightarrow \infty} [\partial y_{t+k} / \partial e_t] = \lim_{k \rightarrow \infty} c_k = 0. \quad (3)$$

The above quantity measures the *persistence* of the shocks hitting the series. Such disturbances can be taken to reflect, in a highly stylised way, the occurrence of demand side shocks, which induce the observed series to deviate temporarily from its long run fundamentals (technological progress, demographic factors, accumulation) driving it along its steady-state path. As in Blanchard and Fischer (1989, p.8), the model can be intended as a canonical form, in which the single stochastic term is taken to represent the action of a plurality of random shocks. This is coherent with the analysis of Granger and Morris (1976).

Secondly, it is possible to conceive of the observed series as generated by the cumulated ever-lasting effects of purely random shocks. This hypothesis has found technical justification in the Box-Jenkins ARIMA modelling framework, where it is recommended that the series be differenced until stationarity is achieved. This means that in the proposed  $ARMA(p,q)$  representation:

$$\rho(L)y_t = \theta(L)e_t \quad (4)$$

the autoregressive polynomial,  $\rho(L)$ , is supposed to have  $d$  unit roots; hence once the series has been differenced  $d$  times,  $\Delta^d y_t$  admits  $ARMA(p^*,q)$  representation with  $p^* = p - d$ . The original series is then called 'difference stationary', or to possess  $d$  unit roots in its autoregressive polynomial.

If attention is restricted to first differencing, as seems to be plausible for many macroeconomic aggregates, then we have a stationary ARMA model for the first differences:

$$\rho^*(L)(\Delta y_t - \delta) = \theta(L)e_t. \quad (5)$$

The series is called integrated of order one. As Beveridge and Nelson (1981) show, any such process can be decomposed into two different random components, the first being a random walk with drift, and the second a stationary zero-mean ARMA process:

$$\begin{aligned} y_t &= y_t^p + y_t^s, \\ \Delta y_t^p &= \delta + [\theta(1)/\rho^*(1)]e_t = \delta + c(1)e_t, \\ y_t^s &= c^*(L)e_t, \\ c(L) &= [\theta(L)/\rho^*(L)] = c^*(L)\Delta + c(1). \end{aligned} \tag{6}$$

The drift arises only if the differenced process has a non-zero mean. The first component can be interpreted as accounting for the growth of the process: its differences deviate randomly from the non-zero expected value given by the drift, with a variance which is  $[c(1)]^2\sigma_e^2$ . The second component gives the transient short run dynamics. In this model shock have an ever-lasting effect on the level of the series given by a non-zero persistence measure:

$$\lim_{k \rightarrow \infty} [\partial y_{t+k} / \partial e_t] = c(1) \neq 0. \tag{7}$$

A synthetic measure of the importance of the non-stationary component  $y_t^p$  is therefore given by the sheer size of  $c(1)$ , which is a non-linear function of the parameters of the AR and MA coefficients.

Note however that both components are by definition generated in terms of the same random disturbance term  $e_t$ , hence the long run and transitory dynamics cannot be considered as distinct, and the shocks affect the variable in a permanent way through their effect on  $y_t^p$ . This is just one of the possible decompositions which can be achieved on a univariate basis. Harvey (1990) presents a

decomposition in terms of different orthogonal processes, which forms the basis of the 'structural time series modelling'.

Models like (6), when applied to variables such as *gnp* or its components, are compatible with a very different explanation of the observed fluctuations, namely that provided by the "real business cycle" theory. See, for example, Plosser (1982), Kydland and Prescott (1982), Prescott (1986), King, Plosser and Rebelo (1988), Campbell (1994). In such a theoretical framework, the role of inter-temporal optimising behaviour of rational agents in labour-leisure and consumption-investment choices is emphasised, and shocks from different sources are allowed to produce permanent effects on the series itself, by modifying its random growth pattern. These shocks are mainly, but not only technological: disturbances can affect tastes and preferences, and can take the form of public sector interventions (see Baxter and King, 1993 or Campbell 1994), or of unexpected changes in the terms of trade (as in Mendoza 1991, or Correia et al. 1995). It is not even necessary that such shocks be given a non-stationary specification to produce permanent effects, given the inter-temporal capital accumulation process. In the short run, the occurrence of disturbances produces temporary effects through the adjustment mechanism, which are intended to be accommodated in the second term of the above decomposition. Moreover such short-run disturbances are part of the optimal reaction of economic agents. For this reason they should be considered as Pareto optimal and not as something the government should aim at offsetting. This is not the only possible explanation of the occurrence of persistent shocks, and its validity is strongly questioned (see Mankiw, 1989). Campbell and Mankiw (1987) regard this persistence as better explained by the presence of nominal rigidities or by multiple equilibria.

In many other fields of economic analysis recent theories imply difference stationarity of the series they purport to explain. In many examples,

*'the presence of a unit root is often a theoretical implication of models which postulate the rational use of information available to economic agents. Examples from economics include various financial market variables, such as future contracts [...], dividends and earnings [...], spot and forward exchange rates [...], and even aggregate variables like real consumption [...] and investment [...].'* (Perron (1988, p.297)

From a more technical viewpoint, in the recent methodological literature it has been emphasised that the correct starting step in econometric modelling is a 'well defined estimated statistical model (Spanos, 1986)' to account for the statistical properties of the series under study. Therefore, it is important to discriminate between the two kinds of non-stationarity at the outset, also in the light of the impact on properties of estimators.

In many studies, e.g. Phillips and Durlauf (1986), Park and Phillips (1988) and (1989), West (1988) and Sims, Stock and Watson (1990) inter alia, properties of estimators have been analysed in regression contexts where some of (or all) the variables involved are integrated. In such cases, estimators may not have asymptotic normal distributions, and they usually do not have. Therefore all the tests that are commonly used in regression analysis have asymptotic distributions which deviate from the usual  $\chi^2$ , and need numerical tabulation.

For all these reasons a host of applied studies have been conducted on a univariate basis, to discriminate between trend and difference stationarity. The first and most influential paper in that respect is Nelson and Plosser (1982), where most of the U.S. macroeconomic variables considered have been found integrated of order one. The technical instruments used there are the usual unit root tests put forward by D.A. Dickey and W.A. Fuller, which are discussed in the following section, after providing an overview of the properties of the *OLS* estimates in the trend stationary and difference stationary cases.

## [2.2] Inference on trend stationary and integrated univariate processes.

Let us consider a simple zero-mean first order autoregressive process:

$$\begin{aligned} y_t &= \rho y_{t-1} + e_t, \\ e_t &\sim i.i.d.(0, \sigma^2). \end{aligned} \quad (8)$$

Clearly, when  $|\rho| < 1$  the model is an extremely simple stationary process, with stationary oscillations around its zero unconditional mean. The dynamic effect of shocks can be described via the moving average representation:

$$\begin{aligned} y_t &= c(L)e_t, \\ c_i &= \rho^i, \quad i = 1, 2, \dots, \infty. \end{aligned} \quad (9)$$

Being given a sample of  $T+1$  observations,  $y_0, y_1, \dots, y_T$ , the simplest way to estimate the unknown parameters of the above process is to use ordinary least squares, which entails maximising the likelihood conditioned on the initial observation  $y_0$ :

$$\hat{\rho} = \frac{\sum_{t=1}^T y_t y_{t-1}}{\sum_{t=1}^T y_{t-1}^2}, \quad \hat{\sigma}^2 = (T-1)^{-1} \sum_{t=1}^T (y_t - \hat{\rho} y_{t-1})^2. \quad (10)$$

It is well known, since the work by Mann and Wald (1943), that the asymptotic distribution of the normalised estimate is normal:

$$T^{1/2}(\hat{\rho} - \rho) \xrightarrow{d} N[0, (1 - \rho^2)]. \quad (11)$$

Therefore, given the sample size, the precision of the estimate of  $\rho$  is an increasing function of  $|\rho|$ , as it is clearly seen in expression (11). The same conclusion can be

drawn from a different viewpoint, which will become relevant in the next chapter. Assuming that the error terms are Gaussian, the conditional log-likelihood function of the model, reads:

$$\log L(\mathbf{b}) = k - T \log \sigma - \frac{1}{2\sigma^2} \sum_{t=1}^T (y_t - \rho y_{t-1})^2, \quad (12)$$

$$\mathbf{b} = \begin{bmatrix} \sigma^2 \\ \rho \end{bmatrix}.$$

The Fisher information matrix  $I_T(\mathbf{b})$  is:

$$I_T(\mathbf{b}) = \left[ -E(\partial^2 \log L / \partial \mathbf{b} \partial \mathbf{b}') \right] = \begin{bmatrix} T/2\sigma^4 & 0 \\ 0 & T/(1-\rho^2) \end{bmatrix} \quad (13)$$

Given the notorious equivalence of the *OLS* and the conditional *ML* estimators, and the usual asymptotic properties of the latter one, the second diagonal element of the information matrix conveys immediately the Mann and Wald (1943) result.

Nevertheless, it is important to stress that in finite samples the shape of the distribution of  $\hat{\rho}$  is very different from normal, being skewed to the left, the skewness being an increasing function of the true unknown value of  $\rho$ .

A generalisation of result (11) holds for the estimation of a stationary *AR* ( $p$ ) process:

$$y_t = \rho_1 y_{t-1} + \rho_2 y_{t-2} + \dots + \rho_p y_{t-p} + \varepsilon_t \quad (14)$$

stating that the scaled *OLS* estimate of  $\rho = [\rho_1, \rho_2, \dots, \rho_p]'$  is asymptotically multivariate Normal:



$$T^{1/2}(\hat{\rho} - \rho) \xrightarrow{d} N[0, \sigma^2 V^{-1}], \quad (15)$$

$$V = \text{var} \begin{bmatrix} y_{t-1} & y_{t-2} & \dots & y_{t-p} \end{bmatrix}$$

Therefore, also for higher order stationary AR processes, inference can be conducted by using asymptotic normality.

Furthermore, if the stationary model is augmented to include a linear time trend:

$$y_t = \alpha + \beta t + \rho_1 y_{t-1} + \rho_2 y_{t-2} + \dots + \rho_p y_{t-p} + \varepsilon_t \quad (16)$$

the scaled OLS estimator of  $\mathbf{b} = [\rho', \alpha, \beta]'$  is asymptotically multivariate Normal:

$$T_T(\hat{\mathbf{b}} - \mathbf{b}) \xrightarrow{d} N[0, \sigma^2 \mathbf{V}^{*-1}], \quad (17)$$

$$\mathbf{V}^* = \begin{bmatrix} \mathbf{V} & 0 & 0 \\ 0 & 1 & 1/2 \\ 0 & 1/2 & 1/3 \end{bmatrix}, \quad \mathbf{T}_T = \begin{bmatrix} T^{1/2} \mathbf{I}_p & 0 & 0 \\ 0 & T^{1/2} & 0 \\ 0 & 0 & T^{3/2} \end{bmatrix}$$

$$V = \text{var} \begin{bmatrix} z_{t-1} & z_{t-2} & \dots & z_{t-p} \end{bmatrix}, \quad z_t = y_t - \alpha - \beta t.$$

Note that the trend term of the deterministic component has a higher rate of convergence than the other coefficients and this circumstance calls for the use of different scaling factors through the apt definition of the scaling matrix  $\mathbf{T}_T$  (see Sims, Stock and Watson 1990).

Given the distributional results described above, it is possible to make standard inference on the parameters of a stationary AR process. Things seem to radically change when dealing with processes with a unit root. In the case the process is difference stationary, radically different properties attain for the parameters estimates. Early studies in this respect are Fuller (1976), Dickey (1976), Dickey and Fuller (1979, 1981). More recently Phillips (1987) has given a more formal description of these properties by making extensive use of the functional central limit theorem as applied to functionals of Brownian motion processes.

Let us consider the simple model (8) where the true value of  $\rho$  is unity. In order to describe the asymptotic properties of the *OLS* estimate of  $\rho$ , it is necessary to define:

$$S_T(r) = \sum_{t=1}^{[Tr]} e_t, \quad r \in [0,1], \quad S_T(r): [0,1] \rightarrow \mathbf{R}, \quad (18)$$

with  $[Tr]$  denoting the integer part of  $Tr$ . A very useful result is what in the literature is termed as "Donsker's theorem", or "functional central limit theorem", or "invariance principle" (see Billingsley 1968), stating that:

$$T^{-1/2} \sigma^{-1} S_T(r) \Rightarrow W(r), \quad (19)$$

where the random variable  $W(r)$  is a Brownian motion process, and  $\Rightarrow$  denotes weak convergence of the associated probability measure, this property is the analogue of convergence in distribution as applied to function spaces. It is possible to see that this result holds also in the presence of less strict requirements for the error terms  $e_t$ , for instance when there is some correlation and time-heterogeneity among them, and this is used in Phillips (1987). Expression (19) states the asymptotic normality of any normalised sample mean obtained by making use of sub-sets of the sample observations. For instance, when  $r=1$ ,  $S_T(1)$  is the normalised sample mean obtained using the full sample, whose asymptotic distribution is clearly  $N(0,1)$ , and it is known that this is indeed the distribution of  $W(1)$ . In other terms (18) is a more general statement of the central limit theorem. A second very useful result is given by the "continuous mapping theorem", (Hall and Heyde, 1980) stating that if  $S_T(r)$  converges to  $S$ , and  $g(S_T(r))$  is a continuous functional, then  $g(S_T(r))$  converges to  $g(S)$ .

Given these results, it is possible to obtain the asymptotic distribution of the normalised *OLS* estimate of  $\rho$ :

$$T(\hat{\rho} - 1) \Rightarrow \frac{(1/2)\{[W(1)]^2 - 1\}}{\int_0^1 [W(r)]^2 dr} \quad (20)$$

It is necessary to stress some features of the distribution above, contrasted to the ones of the analogue asymptotic distribution in the case of stationarity. First of all, in order to have a non-degenerate distribution, it is necessary to scale  $(\hat{\rho} - 1)$  by  $T$  and not by  $T^{1/2}$  as in the stationary case. In other terms  $\hat{\rho}$  is  $O_p(T)$  and not  $O_p(T^{1/2})$ . Secondly, although the finite sample distribution of  $\hat{\rho}$  is skewed both under the stationary and the integrated case, in the latter case the asymptotic distribution of the adequately normalised bias retains its skewness, whereas it is Gaussian in the former case.

On the basis of the continuous mapping theorem, also the asymptotic distribution of the  $t$  statistic for testing the presence of a unit root can be directly obtained:

$$t(\hat{\rho} = 1) = \frac{\hat{\rho} - 1}{\left\{ \hat{\sigma}^2 \left[ \sum_{t=1}^T y_{t-1}^2 \right]^{-1} \right\}^{1/2}} \Rightarrow \frac{(1/2)\{[W(1)]^2 - 1\}}{\left\{ \int_0^1 [W(r)]^2 dr \right\}^{1/2}} \quad (21)$$

The random variables defined in the expressions (20) and (21) can be simulated and therefore it is possible to obtain the desired quantiles. Tables for both statistics are contained in Fuller (1976, p. 371). Together with the asymptotic quantiles, Fuller provides also the quantiles for different finite sample sizes by means of direct numerical simulation of the model under the hypothesis of a unit root. To do that Fuller needs the extra assumption of Gaussian errors.

With the results described above, it is possible to give a solution to the inferential problem of deciding whether or not the simple  $AR(1)$  model (8) has a unit root. In a classical Neyman-Pearson approach, the most immediate way to describe the

hypotheses involved is  $\{H_0: \rho = 1, H_1: \rho < 1\}$ , leaving aside the uninteresting issue about the presence of a explosive root. This is the way taken by Dickey (1976), Fuller (1976), Dickey and Fuller (1979, 1981). Given the asymmetry with which the hypotheses are treated in a classical inferential approach, what becomes relevant is the distribution of the relevant statistics under the null hypothesis of difference stationarity. We have just seen that these distributions are non-standard and call for reference to the tables provided by Fuller.

When it is required to discriminate between different sensible ways of accounting for non-stationarity, the simple  $AR(1)$  model without deterministics is not satisfactory and has to be augmented in two dimensions: include a sensible deterministic part, and consider higher dynamics in the autoregressive representation.

As for the deterministic components, the most straightforward way to achieve this augmentation is to linearly append the desired deterministic component to the model. For this reason, starting from the aforementioned works by Dickey and Fuller, the following two models are commonly specified:

$$y_t = c + \rho y_{t-1} + e_t, \quad (22)$$

$$y_t = c + \beta t + \rho y_{t-1} + e_t. \quad (23)$$

Notice that in the above models the  $c$  and  $\beta$  coefficients have different interpretations depending on whether or not the model is integrated. When the model is stationary, for  $y_t$   $c$  and  $\beta$  clearly define the intercept and the linear trend terms respectively, but when there is a unit root,  $c$  is the linear trend coefficient in model (22), and  $\beta$  is twice the quadratic trend coefficient in model (23). Therefore, the parameterisation used in the two models above can give the possibility of discriminating between trend stationarity and difference stationarity but only by jointly considering a set of statistics to test the following hypotheses:

(1)	$c = 0$	against	$c \neq 0$	in models (22), (23)
(2)	$\beta = 0$	against	$\beta \neq 0$	in model (23)
(3)	$\rho = 1$	against	$\rho < 1$	in models (22), (23)
(4)	$c = 0, \rho = 1$	against	$c \neq 0, \rho < 1$	in model (22)
(5)	$\beta = 0, \rho = 1$	against	$\beta \neq 0, \rho < 1$	in model (23)
(6)	$c = 0, \beta = 0, \rho =$	against	$c \neq 0, \beta \neq 0, \rho < 1$	in model (23)

The asymptotic distributions of these statistics under the integration hypothesis can be obtained using the concept of convergence on function spaces, as in the simplest no-deterministics case. The only complication arises for the presence in the estimated coefficient vectors of terms with different rates of convergence under the integration hypothesis. The main results can be summarised as follows

a) Estimating model (22) one has to distinguish between the circumstance that the "true" value of  $c$  be zero or not. In the former case,  $T(\hat{\rho} - 1)$  and  $t(\hat{\rho} - 1)$  have non standard asymptotic distributions given by functionals of Brownian motions. The test for the joint hypothesis labelled as (4) in the table above involves the use of a Wald "F" test which does not have standard distribution. On the other hand when  $c \neq 0$ , the estimated coefficient are asymptotically Gaussian (see West, 1988), once scaled by means of the matrix  $T_r = \begin{bmatrix} T^{1/2} & 0 \\ 0 & T^{3/2} \end{bmatrix}$ . In this case one can

asymptotically rely on the standard critical values. The reason why this result is obtained is that the presence of a drift term renders  $y_t$  asymptotically dominated by the resulting linear trend. Therefore the regressor  $y_t$  is asymptotically equivalent to a trend, and therefore asymptotic normality of the coefficients attains.

b) Estimating model (23), the only relevant case under the null is when  $\beta = 0$ , otherwise the model will imply non-stationarity around a quadratic trend, which is not conceptually adequate for most economic time series. The statistics being used to test the hypotheses (3), (5) and (6) are all different functionals of a Brownian

motion process. As in the case of the simpler model (8), the finite sample quantiles of the relevant statistics were obtained by Dickey and Fuller (1979, 1981), for different sample sizes, via direct simulation of the dgp with Gaussian disturbances. I turn now to the problem of unit root inference in models with richer dynamics. From an estimation point of view, the most straightforward way to augment the model is to specify an  $AR(p)$  for  $y_t$ , as in Dickey and Fuller (1981). If we consider, for instance, the model with a linear trend we have:

$$\begin{aligned}\rho(L)y_t &= c + \beta t + e_t, \\ \rho(L) &= 1 - \rho_1 L - \rho_2 L^2 - \dots - \rho_p L^p.\end{aligned}\tag{24}$$

This model can be easily reparameterised as:

$$\begin{aligned}\rho^*(L)\Delta y_t &= c + \beta t + \rho y_{t-1} + e_t, \\ \rho(L) &= \rho^*(L)\Delta + \rho, \\ \rho^*(L) &= 1 - \rho_1^* L - \rho_2^* L^2 - \dots - \rho_{p-1}^* L^{p-1}, \\ \rho &= \rho(1).\end{aligned}\tag{25}$$

The unit root occurs when  $\rho = 0$ . The testing strategy consists then in 'augmenting' the model with  $p-1$  lagged differences, and in evaluating the same statistics as in the simpler  $AR(1)$  case above. The resulting tests are known as 'Augmented Dickey Fuller' (*ADF*) tests.

In the presence of any kind of deterministic components in the model and in the dgp, it is possible to show that the limiting distribution of the test statistics coincide with the ones that would be valid for the  $AR(1)$  model with the corresponding deterministic part. The intuition behind this result is that the model is augmented with a set of stationary regressors (the lagged  $\Delta y$ 's) which do not alter the asymptotic distribution of the relevant parameters. In more formal terms this happens because the model can be transformed such that the aptly scaled variance-

covariance matrix of the coefficients is asymptotically block-diagonal (see Banerjee *et al.*, 1993). Therefore the use of the same critical values is asymptotically correct. Model augmentation poses an additional complication, since the validity of the test results depends crucially on the correct identification of the lag order. When a general *ARMA* representation of unknown order is allowed, things get even more complicated, since its approximation via the autoregressive representation could require a very high lag order; Said and Dickey (1984) develop a truncation rule meant to yield consistent results. Hall (1988) considers the problem from a different viewpoint: the simultaneity between regressors and the error term induced by any finite order truncation. He proposes test statistics based on instrumental variable estimation.

A completely different route is suggested in Phillips (1987), Perron and Phillips (1987), Phillips and Perron (1988), and Perron (1988). Their idea rests on Solo (1984), and consists in referring to an *AR*(1) model with the desired deterministic component. The *i.i.d.* hypothesis on the errors is abandoned, and  $e_t$  is considered as an infinite dimensional time dependent and weakly heterogeneous nuisance parameter. The conditions imposed on it are the following ones (see Phillips 1987, p 280):

- a)  $E(e_t) = 0 \forall t$ .
- b)  $\sup_t E(|e_t|^\beta) < \infty$  for some  $\beta > 2$ .
- c)  $\lim_{T \rightarrow \infty} E(T^{-1} S_T^2) = \sigma^2 > 0$ , where  $S_T = \sum_{t=1}^T e_t$ .
- d) The process  $e_t$  is strongly mixing with mixing coefficients  $\alpha_t$  obeying  $\sum_{t=1}^{\infty} \alpha_t^{(1-2/\beta)} < \infty$ .

These conditions imposed on  $e_t$  define a time dependent, weakly heterogeneous process to which extensions of the central limit theorem can be validly applied. The class of processes satisfying these requirements is wide enough to include any stationary *ARMA* or *ARMAX* (with stationary exogenous variables) process.

The first order autoregressive model is estimated by means of *OLS*, and the statistics proposed by Dickey and Fuller are shown to have limiting distributions differing from the corresponding ones attained with *i.i.d.* errors, because of the presence of the nuisance parameter  $\lambda = (\sigma^2 - \sigma_e^2)/2$ , where  $\sigma^2$  is defined in property (c) and  $\sigma_e^2 = \lim_{T \rightarrow \infty} T^{-1} \sum_{t=1}^T E(e_t^2)$ . In fact, when errors are *i.i.d.*  $\sigma^2 \equiv \sigma_e^2$  and  $\lambda \equiv 0$ .

Under weak stationarity for  $e_t$ , the quantity  $\sigma^2$  is  $(2\pi)$  times the spectral density function of  $e_t$  at its origin, and can therefore be non-parametrically estimated in the time domain by means of a finite number of autocovariances. Following Newey and West (1987), smoothing is achieved via a triangular Bartlett window, to ensure the non-negativity of the resulting estimate:

$$\hat{\sigma}^2 = T^{-1} \left( \sum_{t=1}^T \hat{e}_t + 2 \sum_{j=1}^m w(j, m) \sum_{t=j+1}^T \hat{e}_t \hat{e}_{t-j} \right),$$

$$w(j, m) = \max[0, 1 - j / (m + 1)].$$

In this respect, it is necessary to choose the truncation parameter  $m$ , the bandwidth of the estimate. With  $m$  growing with the sample size and defined to be  $o(T^{1/4})$ , consistent estimates of  $\sigma^2$  can be obtained.

The quantity  $\sigma_e^2$  is consistently estimated as  $\hat{s}_e^2 = T^{-1} \sum_{t=2}^T \hat{e}_t^2$ . In this way it is possible to estimate consistently the nuisance parameter  $\lambda$ , and the Dickey-Fuller statistics can be corrected to take it into account. As a result, the available critical values can be validly used.

Unfortunately, many problems arise in the empirical application of the unit root tests so far described. As already pointed out, the operative testing strategy consists in taking jointly into account a plurality of tests, whose outcomes may happen not to be mutually coherent. Moreover, the tests have very low power against specific stationary alternatives. A well-known example is when the



presence of a structural break is envisaged. In this case the above testing procedures tend to over-estimate the order of integration. In Perron (1989) the conclusions of Nelson and Plosser (1982) are endorsed for most of the series considered, on the basis of a different testing procedure allowing for the presence of a structural break. In addition, by means of Monte Carlo simulations, Schwert (1987, 1988) shows that the distributions of the Dickey-Fuller statistics and of their non-parametric corrections have a remarkably slow convergence to their limiting distributions. This fact can induce important differences between nominal and effective sizes in empirical applications. Finally, the way in which the presence of a deterministic trend is accounted for seems deeply unsatisfactory. In fact, as has been pointed out already, the parameters of the deterministic components have different interpretations when the series is stationary and when it is integrated: under the null, the deterministic component being considered is a quadratic trend. Dealing for example with model (23), in order to have a deterministic trend of the same order under both hypotheses, the parameter  $\beta$  should be set equal to zero when  $\rho = 1$ . It is hence redundant under the integration null, inducing a power loss in the statistics based on this model. For this reason, Schmidt and Phillips (1992) describe as 'clumsy' this parameterisation and suggest using the more convenient alternative proposed by Barghava (1986): in the simple  $AR(1)$  case, we can write:

$$x_t = \rho x_{t-1} + e_t, \quad x_t = y_t - \mu - \delta t. \quad (26)$$

Note that the model is non-linear in the parameters and that  $\mu$  disappears when  $\rho = 1$ . In this context no parameter is redundant under either hypothesis: when the series is trend stationary,  $\mu$  and  $\delta$  indicate respectively intercept and slope of the deterministic trend; when the series is integrated,  $\mu$  vanishes and  $\delta$  gives the drift. On the basis of this parameterisation, Schmidt and Phillips develop two LM test statistics, a  $t$ -ratio and a coefficient test, which can be easily corrected to account

for weakly dependent and heterogeneous error terms in the same way as the Dickey and Fuller tests.

Given the difficulties encountered, some authors, like Campbell and Mankiw (1987), propose to reverse completely the testing framework: the series is differenced at the outset; then, embedding stationarity under the null, one should test for 'over-differencing', i.e. for the presence of a unit root in the MA polynomial. This is closely related to the Beveridge-Nelson representation seen in expression (6) above: when a stationary process is differenced, in the Beveridge-Nelson representation we have  $c(1)=0$ .

Unfortunately, in the estimation of MA parameters the so-called 'pile-up' problem emerges: by means of simulation experiments on the basis of a  $MA(1)$  process, Sargan and Barghava (1983, section 5) show that the event that the  $ML$  estimator is equal to one has a positive probability mass even when the 'true' parameter value is not equal to one but close to it. Therefore testing procedures based on the estimated  $MA$  parameters could lead to spuriously detecting over-differencing.

For this reason, other routes should be followed. A possible solution, suggested by Nerlove and Pinto (1984) is to resort to alternative likelihood-based estimation techniques, such as the ones based on frequency-domain approximation of the likelihood function. Another, already in use, consists in focusing in the spectral density function of the differenced series, to check whether it gets to zero at the origin, as it should in case of over-differencing. In this respect, see Ouliaris, Perron and Phillips (1986).

### **[2.3] Some General Considerations About Unit Root Testing in a Classical Framework**

In synthesis, all the classical procedures for unit root testing share some unpleasant features.

(i) The necessity to refer only to asymptotic non standard results, which is given by the particular properties of integrated processes. In finite samples, the distributions of the test statistics are indeed of unknown form, and the only information available in this respect come from Monte Carlo investigations. As we have already stressed, Schwert (1989) demonstrates how the finite sample properties of most unit root tests have radically different properties from their asymptotic counterparts. For this reason, the actual size of tests tend to be substantially different from the nominal one. Schwert (1989) compares the properties of *ADF* and Phillips-Perron tests on data generated as *ARIMA*(0,1,1) processes, finding that the actual size tends to get extremely high when the *dgp* has a large negative *MA* coefficient.

(ii) Generally poor power performances. This is the most difficult aspect relating to the empirical application of these tests, i.e. the evident difficulty of discriminating, in finite samples, between alternative models of non-stationarity which replicate sufficiently well the correlation properties of the series under analysis. This is the so-called "near observational equivalence" described in Sims (1989) and Campbell and Perron (1991). DeJong, Narkervis, Savin and Whiteman (1992) conducted an interesting simulation exercise and report type II errors comparable to ones attained in a coin tossing game.

(iii) Different deterministic components in the *dgp* and in the estimated model imply different asymptotic distributions. This is a source of potential confusion, since it is first necessary to decide which deterministic component to take into consideration. Moreover, when the deterministics are linearly appended, it is

necessary to resort to a battery of different test statistics, whose combined outcome often has a problematic interpretation.

(iv) Asymmetry in the treatment of the hypotheses. In the usual Neyman-Pearson approach, the test procedure assigns different treatment to the two possible wrong outcomes. The probability of falsely discarding the null is fixed for whichever sample size, wherever the probability of rejecting the alternative tends to converge to zero as the sample size increases, in the absence of any model misspecifications.

It is possible to try and solve each one of these problem, just by resorting to Bayesian inferential techniques. First of all, the Bayesian framework does not rely on asymptotic distributions, given that posterior inference is carried out on the basis of the relevant finite sample distributions. Secondly, it is possible to decrease the extent of the "observational equivalence" problem by allowing the researchers to implement explicitly their own personal a priori information about the parametric nature of the data generation process (henceforth *DGP*), in this way increasing the efficiency of the resulting estimate. Moreover, it turns out that different deterministic components do not have any relevant impact on the marginal posterior distributions of the relevant parameters. Finally, we will see that the Bayesian analysis allows one to compare hypotheses on the basis of the corresponding posterior probabilities. The posterior distributions under both the hypotheses are not asymmetric, and the testing is fully "consistent", in that the probability of picking the wrong model goes to zero as the sample size increases. This property is shared also by other Bayesian inferential procedures designed to evaluate hypotheses, such as the one based on the relevant highest posterior density (*HPD*) confidence intervals.

Moreover, since the relevance of the unit root inferential problem is often related to some decisions the applied researcher has to take in the model building process, i.e. difference the series, insert a trend, apply a seasonal frequencies filter, it seems conceptually appealing to be able to provide Bayesian techniques that have

consistently proved to be a valid support to decision making in other fields of human activity.

The more general advantages and disadvantages of the Bayesian techniques also apply to this problem. The advantages, together with the ones already described, are essentially related to the fact that the Bayesian approach is simple, and it constitutes the only logical formalisation of the process of learning. The disadvantages are mainly of a computational kind. Bayesian methods are "labour intensive" techniques.

The main aim of this thesis is to provide fresh examples of the ways in which a Bayesian approach can be validly used to deal with non-stationarity issues, trying to avoid some of the difficulties encountered in the use of classical techniques.

## **Chapter 3: The Bayesian Approach to Time Series Analysis: Problems and Methods.**

### **[3.0] An Overview of the Chapter.**

In this chapter, I discuss the main methodological and implementation issues arising in the application of Bayesian techniques in the analysis of time series models. The chapter is organised as follows: Section [3.1] introduces the concept of prior and posterior distributions. Section [3.2] reviews the Bayesian inferential techniques, with particular emphasis being given to the problem of reporting of the results and of model selection. Section [3.3] explains the general criteria followed in the specification of the prior distribution. Section [3.4] explains the computational problems arising in a Bayesian approach and the solutions that are available to solve them. The use of Markov chain Monte Carlo integration is discussed at length, in order to show its success in freeing the researcher from the conjugate prior straitjacket. Section [3.5] gives some relevant examples of Bayesian studies concerning non-stationary univariate models, and the final section discusses the remarks made by Phillips (1991a) concerning the concept of "ignorance priors" in time series models.

### **[3.1] The General Philosophy of the Bayesian Approach.**

Let us consider a parametric model of the kind:

$$f(y_t, x_t, \epsilon_t; \theta), \theta \in \Theta. \quad (1)$$

where  $y_t$  is a  $(n \times 1)$  vector of dependent variables,  $x_t$  is a  $(k \times 1)$  vector of predetermined and/or exogenous variables,  $e_t$  is vector error term,  $\theta$  is a vector of parameters and  $f(\cdot)$  is a given function of its arguments. In the classical approach, the inferential problem can be summarised as follows: a string of  $T$  observations on  $y_t$  and  $x_t$  is available; the researcher has then to obtain a sensible estimate of the unknown vector of parameters  $\theta$ . This amounts to considering data as the unique source of information.

Most of the time it is possible to write down the joint probability distribution function (henceforth pdf) of the whole sample. When regarded as a function of the parameters, this pdf is called likelihood function. Writing the likelihood function requires making an assumption on the distribution of the error terms, and a certain specification for the  $f(\cdot)$  function. To provide an example, let us assume that  $y_t$  is scalar, that the error terms are *i.i.d.* and  $N(0, \sigma^2)$  (henceforth *N.i.d.*( $0, \sigma^2$ )), and that the model is linear in the parameters:

$$y_t = x_t' \beta + e_t, e_t \sim N(0, \sigma^2), \theta = [\beta', \sigma^2]' \quad (2)$$

Then the likelihood function reads:

$$L(\theta|y) = p(y|\theta) = (2\pi\sigma^2)^{-T/2} \exp [-(1/2\sigma^2) e'e], \quad (3)$$

$$e = [e_1, e_2, \dots, e_T]',$$

In the expression above I have considered the dependence of the likelihood function on the exogenous variables as implicit; for this reason I write  $L(\theta|y)$  and  $p(y|\theta)$ .

The classical inference consists in maximising the likelihood function, in order to obtain an estimate of the parameters. Along with this point estimate, comes an

estimate of the associated uncertainty, which is used to construct confidence intervals and to perform hypothesis testing.

The Bayesian approach radically diverges from the classical one, particularly in the way in which the parameter vector is considered. In the Bayesian analysis, there is no such thing as a "true" unknown value of the parameters. On the contrary, these are considered as random unobservable variables, on which the researcher might have some extra-sample (prior) information. The problem is then how to optimally combine sample and non-sample information. This is accomplished by means of Bayes theorem, or the "principle of inverse probability": the likelihood function is regarded as the probability of the sample data given the parameters ( $p(y|\theta)$ ); the extra-sample information about the parameters is in terms of a "prior" distribution  $p(\theta)$ , assigning probability mass to any subset of  $\Theta$ . Using Bayes' theorem, it is possible to write:

$$p(\theta|y) = p(\theta) p(y|\theta) / \left[ \int p(\theta) p(y|\theta) d\theta \right] = p(\theta) p(y|\theta) / p(y), \quad (4)$$

where the integration at the denominator is performed over  $\Theta$ .

The distribution  $p(\theta|y)$  is called the "joint posterior distribution" of the parameter vector. This pdf measures the uncertainty on the parameters which results after combining all the sources of available information, and constitutes the starting point for conducting inference in such a framework. It is important to note that the posterior pdf, given the observed sample  $y$ , is fully described by the product of the likelihood function and the prior distribution. The denominator of expression (4), insofar it is different from zero, can be interpreted as a normalising constant; therefore it is very common to see in the Bayesian literature:

$$p(\theta|y) \propto p(\theta) p(y|\theta). \quad (5)$$



In this respect, it is possible to give the posterior distribution a very useful interpretation: in expression (5), the likelihood function is being weighted using the prior pdf as the weighting function. Under this point of view, the classical approach corresponds to a special case of the Bayesian analysis: when the prior pdf is diffuse over the parameter space, i.e.  $p(\theta) \propto 1$ , the posterior distribution is proportional to the likelihood function, which is the starting point in the classical inference<sup>1</sup>. To provide a very simple example, imagine one is dealing with the linear regression model (2), with only one regressor  $x_n$  and with  $\sigma^2 = 1$ . The researcher has some prior information about the values of the parameter  $\beta$ , which can be summarised in the following prior distribution:

$$\beta \sim N(\mu_\beta, \sigma_\beta^2). \quad (6)$$

By applying Bayes' theorem, the posterior distribution for  $\beta$  is therefore:

$$p(\beta | y) \propto \exp\{-(1/2)[e'e] - [1/(2\sigma_\beta^2)][\beta - \mu_\beta]^2\}, \quad (7)$$

which is the kernel of a univariate Normal distribution with moments:

$$E(\beta | y) = [x'x + \sigma_\beta^{-2}]^{-1}[x'y + \sigma_\beta^{-2}\mu_\beta], \quad \text{Var}(\beta | y) = [x'x + \sigma_\beta^{-2}]^{-1}. \quad (8)$$

Expression (7) states that, for any value of  $\beta$ , the posterior pdf is given by the product between the density given by the prior distribution and the likelihood

---

<sup>1</sup> For the moment, I overlook the problem of assigning non-informative priors to variance parameters, but this will be appropriately described in Section [3.2].

function evaluated at that point. Notice that an "ignorance", or "diffuse" prior would be the one assigning equal prior weights to all the possible values of  $\beta$ , i.e.  $p(\beta) \propto 1$ . Such a prior is "improper", given that:

$$\int p(\beta) d\beta \rightarrow \infty,$$

i.e. the integral of the pdf would not be finite. In other terms, all the integer order moments of the prior, comprised the one of order zero, do not exist. It is immediate to realise that the ignorance prior corresponds to the limit of the prior pdf (6) when the prior variance  $\sigma_\beta^2$  goes to infinity, i.e. when the "strength" of the prior beliefs about  $\beta$  fades away, equivalently one can think about the inverse variance, which is termed prior precision, going to zero. In such a case the posterior moments become:

$$E(\beta | y) = [x'x]^{-1}[x'y], \text{ Var}(\beta | y) = [x'x]^{-1}. \quad (9)$$

These quantities correspond to the usual maximum likelihood estimate of  $\beta$  and of its variance: this is not surprising given that the posterior distribution coincides with the likelihood function. This very simple result holds also in models with more than one regressor and more than one equation.

The way in which the likelihood function and the prior information are combined can be interpreted also under a different viewpoint, which does not require distributional assumptions on the model. With respect to a simple linear regression model:

$$y = x\beta + e, E(e) = 0, \text{ Var}(ee') = \sigma^2 I_n, \quad (10)$$

one can think of having prior information about  $q$  linear combination of the parameters in the form:

$$\mathbf{R}' \boldsymbol{\beta} = \mathbf{d} + \mathbf{e}_0, \quad E(\mathbf{e}_0) = 0, \quad \text{Var}(\mathbf{e}_0 \mathbf{e}_0') = \sigma_p^2 \mathbf{I}_T. \quad (11)$$

This formulation differs from the usual linear constraints in that the extra-sample information about  $\boldsymbol{\beta}$  is subject to error, which implies prior uncertainty. Considering extra-sample information as an additional  $q$  observations leads to a *GLS* mixed estimator

$$\begin{aligned} \tilde{\boldsymbol{\beta}} &= [\sigma^{-2} \mathbf{x}'\mathbf{x} + \sigma_p^{-2} \mathbf{R}\mathbf{R}']^{-1} [\sigma^{-2} \mathbf{x}'\mathbf{y} + \sigma_p^{-2} \mathbf{R}'\mathbf{d}], \\ \text{var}(\tilde{\boldsymbol{\beta}}) &= [\sigma^{-2} \mathbf{x}'\mathbf{x} + \sigma_p^{-2} \mathbf{R}\mathbf{R}']^{-1}. \end{aligned} \quad (12)$$

This is the well-known Theil-Goldberger (1961) mixed estimator. Adding distributional assumptions on  $\mathbf{e}$  and  $\mathbf{e}_0$  delivers the Bayesian posterior pdf.

Having combined sample evidence with non-sample information, the result is the joint posterior pdf of the whole set of parameters of the model, namely  $p(\boldsymbol{\theta}|\mathbf{y})$ . This joint distribution is the starting point in order to address precise questions concerning the model.

### [3.2] Bayesian Inferential Techniques.

In the vast majority of the applications the researcher is typically interested in, inference is conducted only on a strict subset of the whole parameter set of the model. For this reason, the joint posterior distribution is usually marginalised with respect to the parameters the researcher is not interested in. In more formal terms,

if  $\theta = [\theta_1', \theta_2']'$  is the vector of the parameters of the model, and there is interest only in its subset  $\theta_1$ , it is reasonable to work with the marginalised posterior distribution,

$$p(\theta_1|y) = \int p(\theta|y) d\theta_2 \quad (13)$$

This is conceptually a very straightforward step, but, as shown in Section [3.4] it might involve some complications, given that analytical integration is not always feasible.

Another important issue is connected to the problem of reporting. The final result is the marginal posterior pdf for the parameters of interest. The problem of synthesising this posterior distribution for reporting reasons consists in choosing a small number of statistics to summarise the features of the posterior pdf. These can be measures of central tendency, dispersion, skewness, kurtosis, dependence. As Zellner (1971, Section 2.5) proposes, the choice of such synthesis measures is optimally formalised in accordance to the expected utility hypothesis. Calling  $\theta$  the vector of parameters of interest, and  $p(\theta|y)$  the associated posterior pdf, a loss function is specified:  $L(\theta, \hat{\theta}(y))$ , measuring the loss associated in not knowing  $\theta$  and measuring it with a point estimate  $\hat{\theta}(y)$ . The loss function has to be considered a random variable, since  $\theta$  itself is a random variable. The optimal choice of  $\hat{\theta}(y)$  is made by minimising the expected value of the loss function, where the expectation is evaluated on the basis of the posterior pdf of  $\theta$ :

$$\min_{\hat{\theta}(y)} E(L(\theta, \hat{\theta}(y))|y) = \min_{\hat{\theta}(y)} \int L(\theta, \hat{\theta}(y)) p(\theta|y) d\theta. \quad (14)$$

An illuminating example is given in Zellner (1971, p. 24) where it is shown that with a quadratic loss function of the kind:

$$L(\theta, \hat{\theta}(y)) = (\theta - \hat{\theta}(y))' C (\theta - \hat{\theta}(y)), \quad (15)$$

where  $C$  is any positive definite non random matrix, the expected loss is minimised by choosing  $\hat{\theta}(y) = E(\theta|y)$ , i.e. the posterior expectation of the parameter vector

As for the prediction problem in the Bayesian framework, this is also cast in terms of posterior probability. If one has to forecast the vector  $y_k^* = [y_{t+1}, y_{t+2}, \dots, y_{t+k}]$ , i.e. the next  $k$  future values of  $y_t$ , it is possible to define:

$$p(y_k^*|y) = \int p(y_k^*|y, \theta) p(\theta|y) d\theta, \quad (16)$$

where  $p(y^*|y)$  is the posterior pdf of the forecast and  $p(y^*|y, \theta)$  is the likelihood of the future observations conditioned upon sample evidence and knowledge of the parameters. It is immediate to see that the posterior pdf of the future observations is given by their likelihood weighted by the posterior pdf of the parameters. This distribution is a complete measure of the uncertainty associated with the forecasts. It can be synthesised with a point estimate only by choosing a risk function to be minimised.

Attention should also be given to the way in which hypotheses are compared in a Bayesian framework. In Bayesian analysis hypotheses are compared on the basis of the 'posterior odds ratio' (henceforth *POR*, see Zellner, 1971, or Leamer, 1978), that is the ratio between the posterior probabilities associated with the different hypotheses under scrutiny. Following Zellner (1971, ch.8), we consider two hypotheses concerning the parameter space  $\Theta$  associated with a certain model:

$$H_0 \text{ such that } p(H_0, \Theta) = p(H_0) p(\Theta|H_0) = p(H_0) p(\Theta), \quad (17)$$

$$H_1 \text{ such that } p(H_1, \Theta) = p(H_1) p(\Theta|H_1) = p(H_1) p(\Phi),$$

where  $\theta$  and  $\phi$  indicate particular subsets of  $\Theta$ . It is therefore also possible to think of the two hypotheses as implying completely different parameterizations on the model. The two hypotheses have associated prior probabilities:

$$p(H_0)=p_0, p(H_1) = 1 - p_0. \quad (18)$$

Starting from the prior probabilities, the posterior probabilities are obtained by applying Bayes' theorem:

$$p(H_0) \propto p_0 \int p(\theta) p(\text{data} | \theta) d\theta, \quad (19)$$

$$p(H_1) \propto (1-p_0) \int p(\phi) p(\text{data} | \phi) d\phi,$$

where  $p(\text{data} | \phi)$  and  $p(\text{data} | \theta)$  are the likelihoods associated respectively to  $\phi$  and  $\theta$ . The posterior odds ratio is then easily obtained as follows:

$$K_1 = \frac{p(H_0|y)}{p(H_1|y)} = K_0 \frac{\int p(\theta) p(\theta|y) d\theta}{\int p(\phi) p(\phi|y) d\phi} \quad (20)$$

The second term above is known as the 'Bayes factor'; it gives the way in which data evidence is allowed to modify the prior odds ratio  $K_0 = p_0/(1-p_0)$ , i.e. the researcher's prior assessment of the relative plausibility of the hypotheses.

Note that, unlike in the classical Neyman-Pearson hypothesis testing framework, the two hypotheses considered are given a completely symmetric treatment. These hypotheses can take a very wide range of forms, everything depending on how  $\phi$  and  $\theta$  are specified. Some caution is required in contexts where a sharp point hypothesis is compared to a composite one: in such circumstances, it is necessary

to specify, under the point hypothesis, a prior distribution assigning positive probability mass to the corresponding submanifold of  $\Theta$ , say  $\theta$ , while  $\phi$  represents the region of parameter space which is identified by the composite hypothesis. In order to illustrate this point, it might be useful to resort to a simple example. In the simple linear regression model:

$$y_i = \alpha + \beta x_i + e_i, e_i \sim N.i.d.(0, \sigma^2), \quad (21)$$

suppose that the two hypothesis being compared are:

$$H_0: \beta = 0 \text{ vs. } H_1: \beta \neq 0 \quad (22)$$

Clearly  $H_0$  is a sharp point hypothesis, whereas  $H_1$  is a composite alternative. In order to give  $H_0$  a fair chance, it is necessary to specify a prior odds ratio different from zero, i.e.  $p_0 \neq 0$ .

It is important to stress that the use of *POR*'s as a model choice instrument needs a very cautious approach to the specification of priors. First of all, an important caveat relates to the use of improper priors. Only parameters that have a symmetrical role under the two hypotheses can be given an improper prior. If, for instance, we are comparing a point hypothesis with an interval one, the parameters constrained to a point value under  $H_0$  cannot be assigned an improper prior, otherwise the posterior odds ratio would go to infinity as the sample size increases irrespective of any sample evidence against  $H_0$ . This can be easily seen by making a simple example: suppose  $\mathbf{x}$  is a  $(T \times 1)$  vector of  $N.i.d.(\mu, 1)$  draws, and that  $H_0: \mu = 0$ , whereas  $H_1: \mu \neq 0$ . If one specifies

$$p(\mu | H_1) \propto 1, \quad (23)$$

i.e. an improper prior for  $\mu$ , the posterior pdf of  $\mu$  under the alternative is simply the likelihood function:

$$p(\mu | y, H_1) \propto \exp \{ -(1/2)(\mathbf{x} - \mu \mathbf{i})'(\mathbf{x} - \mu \mathbf{i}) \}, \quad (24)$$

where  $\mathbf{i}$  is a  $(T \times 1)$  vector of ones. The posterior odds ratio is hence

$$K_1 = K_0 \frac{(2\pi)^{-1} \exp \left\{ -\frac{1}{2} \mathbf{x}' \mathbf{x} \right\}}{T^{-1/2} \exp \left\{ -\frac{1}{2} \mathbf{x}' \mathbf{M}(\mathbf{i}) \mathbf{x} \right\}} = K_0 \frac{T^{1/2} \exp \left\{ \frac{T}{2} (\bar{x})^2 \right\}}{(2\pi)}, \quad (25)$$

$$\mathbf{M}(\mathbf{i}) = \mathbf{I}_T - \mathbf{i}(\mathbf{i}'\mathbf{i})^{-1}\mathbf{i}', \quad \bar{x} = \mathbf{x}'\mathbf{i}(\mathbf{i}'\mathbf{i})^{-1} = T^{-1} \sum_{t=1}^T x_t.$$

which diverges to infinity even when  $\mu \neq 0$ . Heuristically, this happens because the denominator of the Bayes' factor is obtained as the likelihood function weighted by the prior. As the sample size increases, the likelihood function gets more and more concentrated, but the narrower and narrower range of high likelihood values are averaged together with the parameter values in the tails of the likelihood with equal weights given by the flat improper prior. Therefore the denominator always goes to zero.

Secondly, also the use of flat proper priors could have the same effect. In fact, when using a proper uniform prior for  $\beta$  of the kind:

$$p(\beta) \propto [b - a]^{-1}, \quad a < b, \quad a \in \mathbf{R}, \quad b \in \mathbf{R}, \quad (26)$$

with "too wide" an interval  $[b - a]$ , it is clear that the denominator of the Bayes' factor will tend to zero because a range of values of  $\beta$  with very low likelihood will be averaged with equal weights.



Once the *POR* has been evaluated, it cannot *per se* give any guidance to decisions concerning accepting or rejecting  $H_0$  or  $H_1$ . A decision making criterion can be formulated only by specifying explicitly a loss function and taking decisions as to minimize expected loss on the basis of posterior probabilities. The loss function is taken to measure the disutility associated with 'wrong' decisions (accepting  $H_0$  when  $H_1$  is true or viceversa). If a symmetric loss function is chosen, i.e. the two wrong decisions are associated with equal costs, not surprisingly expected loss is minimized by choosing  $H_0$  ( $H_1$ ) when  $K_1$  is greater (less) than one.

Another important aspect of Bayesian inference is the construction of confidence intervals. In the classical inferential setting, interval estimation is very important because it is an efficient way to synthesise the uncertainty associated with point estimates. In most economic applications, it is possible to make use of the asymptotic normality of the estimates, and to construct asymptotically valid confidence intervals.

In the Bayesian framework, the complete measure of the uncertainty connected to the estimation of the parameters is their joint posterior pdf,  $p(\theta|y)$ . The marginal distributions associated with the single parameters need not be of a known type. In order to give synthetic measures of the posterior uncertainty about, say,  $\theta_i$ , it is possible to evaluate the quantiles  $q_\alpha$  of its posterior pdf:

$$\int_{-\infty}^{q_\alpha} p(\theta_i|y) d\theta_i = \alpha, \quad \alpha \in [0,1]. \quad (27)$$

On the basis of the quantiles, one can then obtain interval estimations based on posterior densities. The quantiles can be used for the construction of highest posterior density (*HPD*) confidence intervals: the  $(1-\alpha)\%$  *HPD* interval is defined as the shortest interval  $[a, b]$  such that:

$$\int_b^b p(\theta_i | \mathbf{y}) d\theta_i = 1 - \alpha, \alpha \in [0,1] \quad (28)$$

In other words, it is the interval associated with the  $(1-\alpha)\%$  probability mass of the posterior pdf  $p(\theta_i | \mathbf{y})$ , collecting the points with the highest posterior densities.

*HPD* intervals can also be a decision criterion in the comparison of hypotheses. For instance, comparing  $H_0: \beta = \beta_0$  and  $H_1: \beta \neq \beta_0$ , it is possible to obtain the posterior probability of  $\beta$ , and use it to construct a *HPD* interval at the desired confidence level and check whether such an interval contains the value  $\beta_0$  or not. In case it does, the preference of the researcher will be for  $H_0$ , and viceversa when  $\beta_0$  falls out of the interval. Notice that this approach does not require such a careful specification of prior distributions as does the one based on posterior odds ratios. Nevertheless, it delivers a fully consistent decision criterion.

### [3.3] The specification of the priors.

The aim of the prior pdf specification is fully to reflect the extra sample information the researcher might want to combine with the sample evidence. The first issue in this discussion is to describe ways in which a situation of prior ignorance should be accommodated. The main contribution in this respect is Jeffreys (1961) and therefore the ignorance priors are also termed "Jeffreys priors". If the inferential problem relates to an unknown parameter  $\beta \in \mathbf{R}$ , a sensible way to reflect absolute prior uncertainty about it could be to specify:

$$p(\beta) d\beta \propto d\beta, \quad (29)$$

which is a flat improper prior. The reason why this pdf is called improper is that it does not integrate to any finite constant, i.e.  $\int p(\beta) d\beta = \infty$ . The rationale behind this

choice is that in this way a situation of complete ignorance is reflected: given the non-overlapping intervals  $[a, b]$  and  $[c, d]$  the ratio of the prior probabilities attached to the events  $\beta \in [a, b]$  and  $\beta \in [c, d]$  is indeterminate.

When inference is drawn on a parameter, say  $\sigma$ , defined on  $\mathbf{R}^+$ , the customary prior distribution reflecting absolute ignorance is obtained by defining  $\theta = \log \sigma$ , and applying on  $\theta \in \mathbf{R}$  the improper flat prior  $p(\theta) d\theta \propto d\theta$ , which implies for  $\sigma$ :

$$p(\sigma) d\sigma \propto d\sigma / \sigma. \quad (30)$$

The prior pdf described in expression (30) is again improper since  $\int p(\sigma) d\sigma \rightarrow \infty$ , but it is nevertheless convenient, because it has some desirable properties. First, the ratio of the prior probabilities attached to  $\sigma \in [0, a]$  and  $\sigma \in (a, \infty)$  is again indeterminate. Second, the prior distribution (30) is invariant to reparameterisations of the kind  $\eta = \sigma^a$ , in the sense that probabilistic statements on  $\sigma$  are consistent with the corresponding ones made on  $\eta$ .

In more general terms, Jeffreys suggested a general rule in order to specify complete ignorance priors. The rule is as follows. Let  $\theta$  be the complete vector of all parameters of the model. The complete ignorance prior pdf is:

$$p(\theta) d\theta \propto |\mathbf{I}(\theta)|^{1/2} d\theta, \quad (31)$$

where  $\mathbf{I}(\theta) = -E (\partial^2 \log L / \partial \theta \partial \theta')$  is the information matrix of the model. The adoption of this rule leads to a prior pdf which enjoys some invariance properties, as described in Zellner (1971, Appendix to Chapter 2). Notice that in the case of a Normal linear regression model of the kind:

$$\mathbf{y} = \mathbf{X} \beta + \mathbf{e}, \mathbf{e} \sim N(0, \sigma^2 \mathbf{I}_T), \quad (32)$$

defining  $\theta = [\beta', \sigma']$ , the prior pdf obtained through the information matrix is:

$$p(\theta) d\theta \propto |X'X|^{1/2} \sigma^{-1} d\theta, \quad (33)$$

which corresponds exactly to the priors specified in expressions (29) and (30). This holds in the linear regression model conditioned on exogenous regressors. On the other hand, as we will see in the Section [3.6], the prior pdf constructed on the basis of the information matrix has radically different features in any linear time series model.

When the researcher wants to reflect some prior information, other kinds of prior should be specified. An analytically convenient strategy is to resort to "natural conjugate" prior distributions (the definition is contained in Raiffa and Schlaifer, 1961. For a quick discussion see Zellner, 1971, *p.* 21). These priors are such that the posterior pdf they generate have the same analytical form as the likelihood function. In this way they give the possibility to implement prior beliefs in a mathematically convenient way. The concept of natural conjugate prior is intimately related to that of sufficient statistics. If we have a model parameterised in terms of the vector  $\theta$ , a set of  $k$  sufficient statistics  $t = [t_1, t_2, \dots, t_k]'$  exists when the likelihood function of the  $(T \times 1)$  data vector  $y$  can be factorised as follows:

$$p(y|\theta, T) = f_1(t|\theta, T) f_2(y); \quad (34)$$

In this context, a natural conjugate prior distribution for  $\theta$  is any density function with the same functional form as  $f_1(t|\theta)$ :

$$p(\theta | t_0, T_0) \quad (35)$$

In this expression the argument of  $p(\cdot)$  is the vector of parameters  $\theta$ , whereas  $t_0, T_0$  are hyperparameters intended to reflect prior beliefs.

Some examples are useful to understand how conjugate priors are specified. Let us consider the linear model (32). It is easy to see that the likelihood function can be factorised as in expression (34), where:

$$f_1(\mathbf{t}|\boldsymbol{\theta}, T) = \sigma^{-T} \exp\left\{-\frac{\mathbf{y}'\mathbf{M}(\mathbf{X})\mathbf{y}}{2\sigma^2}\right\} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})' \mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right\}, \quad (36)$$

and the natural conjugate prior distribution has the form:

$$p(\boldsymbol{\beta}, \sigma) = p(\boldsymbol{\beta}|\sigma) p(\sigma), \quad (37)$$

$$p(\boldsymbol{\beta}|\sigma) \propto \sigma^{-k} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)' \mathbf{C}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right\},$$

$$\text{i.e. } (\boldsymbol{\beta}|\sigma) \sim \mathbf{N}(\boldsymbol{\beta}_0, \sigma^2 \mathbf{C}_0), \quad (38)$$

$$p(\sigma) \propto \sigma^{-(\nu_0+1)} \exp\left\{-\frac{\nu_0 s_0^2}{2\sigma^2}\right\}, \text{ i.e. } \sigma \sim \text{IG}(\nu_0, s_0). \quad (39)$$

The natural conjugate prior is Inverted Gamma - Normal. The hyperparameters  $\boldsymbol{\beta}_0$ ,  $\mathbf{C}_0$ ,  $\nu_0$ , and  $s_0$  are specified as to reflect prior information. When  $\mathbf{C}_0^{-1} = [\mathbf{0}]$  and  $\nu_0 = 0$ , we obtain Jeffreys' ignorance prior pdf.

For any values of these hyperparameters, the joint posterior pdf is analytically tractable. Straightforward analytical integration shows that the marginal posterior pdf of  $\boldsymbol{\beta}$  is multivariate Student- $t$ , with location  $\tilde{\boldsymbol{\beta}} = [\mathbf{X}'\mathbf{X} + \mathbf{C}_0^{-1}]^{-1} [\mathbf{X}'\mathbf{y} + \mathbf{C}_0^{-1}\boldsymbol{\beta}_0]$ , scale parameter  $\tilde{\sigma}^2 [\mathbf{X}'\mathbf{X} + \mathbf{C}_0^{-1}]^{-1}$ , with  $\tilde{\sigma}^2 = \{\nu_0 s_0^2 + \mathbf{y}'\mathbf{y} + \boldsymbol{\beta}_0' \mathbf{C}_0^{-1} \boldsymbol{\beta}_0 + \tilde{\boldsymbol{\beta}}' [\mathbf{X}'\mathbf{X} + \mathbf{C}_0^{-1}] \tilde{\boldsymbol{\beta}}\} + (T + \nu_0)$ , and degrees of freedom equal to  $T + \nu_0$ . This is very convenient, since the Student- $t$  multivariate joint distribution produces Student- $t$  marginal pdf's for any subset of the elements of  $\boldsymbol{\beta}$ . The marginal posterior pdf for  $\sigma$  is Inverted Gamma. In a multivariate context, the natural conjugate prior analysis leads to some difficulties. Let us consider a multivariate regression model of the kind:

$$\mathbf{Y} = \mathbf{X} \mathbf{B} + \mathbf{E}, \text{vec}(\mathbf{E}) \sim \mathbf{N}(0, \Sigma \otimes \mathbf{I}_T), \quad (40)$$

where  $\mathbf{Y}$  is the  $(T \times g)$  matrix collecting the  $T$  observations on  $g$  endogenous variables;  $\mathbf{X}$  is a  $(T \times k)$  matrix of exogenous variables, and  $\mathbf{E}$  is the  $(T \times g)$  matrix of disturbances. The likelihood function is

$$L \propto |\Sigma|^{-T/2} \exp\left\{-(1/2)\text{tr}[\mathbf{E}'\mathbf{E}\Sigma^{-1}]\right\}. \quad (41)$$

The likelihood function can be factorised as in expression (34), where:

$$f_1(\mathbf{t}|\theta, T) = |\Sigma|^{-T/2} \exp\left\{-\frac{1}{2}\text{tr}\left[\mathbf{Y}'\mathbf{M}(\mathbf{X})\mathbf{Y}\Sigma^{-1}\right]\right\} \times \exp\left\{-\frac{1}{2}\text{tr}\left[(\mathbf{B}-\hat{\mathbf{B}})'\mathbf{X}'\mathbf{X}(\mathbf{B}-\hat{\mathbf{B}})\Sigma^{-1}\right]\right\}, \quad (42)$$

and therefore the natural conjugate prior distribution has the following form:

$$p(\mathbf{B}, \Sigma) = p(\mathbf{B}|\Sigma) p(\Sigma), \quad (43)$$

$$p(\mathbf{B}|\Sigma) \propto |\Sigma|^{k/2} \exp\left\{-\frac{1}{2}\text{tr}\left[(\mathbf{B}-\mathbf{B}_0)'\mathbf{C}_0(\mathbf{B}-\mathbf{B}_0)\Sigma^{-1}\right]\right\}, \quad (44)$$

$$p(\Sigma) \propto |\Sigma|^{-(\nu_0+g+1)/2} \exp\left\{-\frac{1}{2}\text{tr}[\mathbf{S}_0\Sigma^{-1}]\right\}, \quad (45)$$

i.e.  $(\mathbf{B}|\Sigma) \sim \text{MN}(\mathbf{B}_0, \Sigma \otimes \mathbf{C}_0)$  and  $\Sigma \sim \text{IW}(\nu_0, \mathbf{S}_0)$ .

The natural conjugate prior is Inverted Wishart-Matricvariate Normal. The hyperparameters  $\mathbf{B}_0$ ,  $\mathbf{C}_0$ ,  $\nu_0$ , and  $\mathbf{S}_0$  reflect prior information. When set to the values  $\mathbf{C}_0^{-1} = [0]$ ,  $\nu_0 = 0$  and  $\mathbf{S}_0 = [0]$ , the hyperparameters define the multivariate Jeffreys ignorance prior pdf, i.e. a diffuse non informative prior on  $\mathbf{B}$ ; in this case, the non-informative prior on  $\Sigma$  is:

$$p(\Sigma) \propto |\Sigma|^{-(g+1)/2}. \quad (46)$$

Using a natural conjugate prior, we obtain the following marginal posterior pdf for  $\mathbf{B}$ :

$$\begin{aligned} p(\mathbf{B}|\mathbf{y}) &\propto |\bar{\mathbf{S}} + (\mathbf{B} - \bar{\mathbf{B}})'(\mathbf{X}'\mathbf{X} + \mathbf{C}_0^{-1})(\mathbf{B} - \bar{\mathbf{B}})|^{-(T+\nu_0)/2}, \\ \bar{\mathbf{S}} &= [\nu_0 \mathbf{S}_0 + \mathbf{B}_0' \mathbf{C}_0^{-1} \mathbf{B}_0 + \mathbf{Y}'\mathbf{Y} + \bar{\mathbf{B}}'(\mathbf{X}'\mathbf{X} + \mathbf{C}_0^{-1})\bar{\mathbf{B}}] / (T + \nu_0), \\ \bar{\mathbf{B}} &= [\mathbf{X}'\mathbf{X} + \mathbf{C}_0^{-1}]^{-1} [\mathbf{X}'\mathbf{Y} + \mathbf{C}_0^{-1} \mathbf{B}_0]. \end{aligned} \quad (47)$$

The posterior pdf of  $\mathbf{B}$  is therefore Matricvariate Student- $t$ , implying that the marginal and conditional posterior pdf's of single rows and columns of  $\mathbf{B}$  are multivariate Student- $t$ . The unpleasant feature of the natural conjugate analysis is that  $\text{vec}(\mathbf{B})$  has posterior variance-covariance matrix equal to  $\bar{\mathbf{S}} \otimes [\mathbf{X}'\mathbf{X} + \mathbf{C}_0^{-1}]^{-1}$ .

The structure of this matrix implies that the ratio of the variances between any two coefficients in equation  $i$  is bound to be the same as the ratio of the corresponding coefficients in any other equation of the system. This has prompted another kind of analysis, the so-called "extended natural conjugate analysis (see Dreze and Richard, 1983, pp. 541 and ff.), based upon a Multivariate Normal prior for  $\beta = \text{vec}(\mathbf{B})$ , where the prior variance-covariance matrix does not have a tensor product structure as  $\Sigma \otimes \mathbf{C}_0^{-1}$ , as the covariance matrix associated with the natural conjugate prior pdf. The resulting posterior distribution is not as easily tractable as the one produced by natural conjugate analysis and calls for a numerical or approximation procedure.

In synthesis, in order to reflect prior information a class of analytically very convenient prior distributions are available, the natural conjugate prior pdf's. Sometimes, though, it could be necessary to resort to other prior pdf's. This could

happen when conjugate priors are not available in the particular problem being treated, or when the natural conjugate prior would generate unwanted features in the posterior pdf. In these cases, it is necessary to be ready to deal with non easily tractable posterior pdf's. The techniques to handle these situations are surveyed in the following section.

### **[3.4] Computational Problems and Technical Solutions**

The main technical problem met by the Bayesian researcher is how to marginalise the joint posterior pdf, in order to focus on the marginal posterior distribution of a subset of the parameters. In other words, if one supposes that  $\theta = [\theta_1', \theta_2']'$  is the complete parameter vector and  $\theta_1$  is the subset of parameters of interest, the problem is to compute

$$p(\theta_1|y) = \int p(\theta|y) d\theta_2 \quad (48)$$

In some cases the joint posterior pdf might not allow analytical integration, and some other solutions are necessary. The solutions are mainly of three types: (a) Approximation of the posterior pdf. (b) Numerical integration. (c) Monte Carlo integration based on numerical simulation.

I describe each of these solutions in the next subsections:



### [3.4.a] Approximations.

The use of expansions of pdf's is long known, especially in the literature dealing with large sample properties of estimators. When the posterior pdf is not of any tractable form, it could be useful to approximate it by means of a tractable one. The most successful approximation is the one based on the second order Taylor expansion of the log pdf around its modal value  $\hat{\theta}$ :

$$\log p(\theta|y) \approx N[\hat{\theta}, H(\hat{\theta}|y)], H(\hat{\theta}|y) = \{-[\partial^2 \log p(\theta|y)/\partial\theta\partial\theta']\}_{\theta=\hat{\theta}}^{-1}. \quad (49)$$

All that it is required to know is the modal value of the distribution. The posterior expectation is approximated by the mode, and the posterior variance covariance matrix by the negative of the Hessian computed at the mode. As Koop (1994) remarks, this entails resorting to the asymptotic normality of maximum likelihood estimates in order to do inference in a classical framework. The approximation is sometimes very bad and completely unreliable when the posterior distribution is not symmetric or multimodal. Moreover, in many occasions the interest of the researcher is in non-linear functions of the parameters; in such cases a Normal approximation of the posterior pdf is pretty pointless.

Another popular approximation is the one used by Phillips (1983) Tierney and Kadane (1986), based on the Laplace approximations of multivariate integrals. If one has to evaluate the posterior expectation of a function of the parameters  $f(\theta)$ , and this is not analytically feasible, the Laplace approximation of the required integral is:

$$\begin{aligned} \int f(\theta) p(\theta|y) d\theta &\approx [H^*/|H|]^{1/2} \exp[L^*(\hat{\theta}^*) - L(\hat{\theta})], \\ L^*(\theta) &= \log [f(\theta) p(\theta) p(y|\theta)], L(\theta) = \log [p(\theta) p(y|\theta)], \\ \hat{\theta}^* &= \operatorname{argmax} L^*, H^* = \{-[\partial^2 L^*/\partial\theta\partial\theta']\}_{\theta=\hat{\theta}^*}^{-1}. \end{aligned} \quad (50)$$

$$\hat{\theta} = \operatorname{argmax} L, \mathbf{H} = \{-[\partial^2 L / \partial \theta \partial \theta']_{\theta = \hat{\theta}}\}^{-1}.$$

All that is needed to use the Laplace approximation is basically optimisation of the functions  $L^*$  and  $L$ .

The drawback of this approximation is that in specific applications it is very difficult to gauge its accuracy. It might be a desired feature of the procedure that the approximation errors tend to vanish as the sample size increases, but very little can be said in terms of the approximation errors obtaining in finite samples. Moreover the approximation might be reasonable in certain regions of the parameter space, and completely unsatisfactory in other regions. In the absence of any information about the properties of the distribution being approximated, the researcher can obtain very misleading results.

#### **[3.4.b] Numerical integration.**

Numerical integration is theoretically always feasible, and different algorithms are nowadays widely available. Typically the Simpson trapezoidal quadrature and the Gauss-Legendre quadrature are the most used algorithms. The technique is very easy when integration has to be performed over a single dimension, or at most two or three. The big drawback of this technique emerges when integration has to be performed over many dimensions. The number of point evaluations of the joint posterior pdf rapidly explodes rendering the technique rather cumbersome. Moreover, in most of the applications, very little is known about the accuracy of the computations.

### [3.4.c] Monte Carlo Integration and Importance Sampling.

This subsection describes how Monte Carlo integration can be used to deal with intractable posterior pdf's. The main references in this respect are Hammersley and Handscomb (1964), Kloek and van Dijk (1978), Ripley (1987) and Geweke (1989). Monte Carlo integration is based on a very simple principle.

Suppose that the joint posterior pdf  $p(\theta|y)$ , although of analytically intractable form, is such that it is possible to generate *i.i.d.* draws from it. In this case the posterior expectation of any function of the parameters could be estimated at the desired level of accuracy in the following way.

Define  $\theta^{(i)}$ ,  $i = 1, 2, \dots, N$  as the  $i^{\text{th}}$  element of a sequence of *i.i.d.* draws from  $p(\theta|y)$ , the joint posterior distribution. Assume that: (a) The posterior pdf is supposed to be proper, i.e.  $\int p(\theta) p(y|\theta) d\theta < \infty$ , (b) the posterior expectation and variance of  $f(\theta)$ , defined respectively as:

$$E[f(\theta)|y] = \int f(\theta)p(\theta|y) d\theta, \text{Var}[f(\theta)|y] = \int [f(\theta) - E[f(\theta)|y]]^2 p(\theta|y) d\theta, \quad (51)$$

are assumed to exist. Under these conditions, this posterior expectation can be estimated as:

$$\hat{E}_T[f(\theta)|y] = N^{-1} \sum_{i=1}^N f(\theta^{(i)}). \quad (52)$$

The properties of this estimator are very easy to obtain. The strong law of large numbers ensures that  $\hat{E}_T[f(\theta)|y]$  converges to  $E[f(\theta)|y]$  as  $N \rightarrow \infty$ . If it is desired to increase the accuracy of the estimate, it is just necessary to increase the number of draws from the posterior distribution of  $\theta$ . This property is often referred to as *simulation consistency*.

It is also possible to obtain an estimate of the standard error of  $\hat{E}_T[f(\theta)|y]$  as:

$$\hat{\sigma}_N = N^{-1/2} \left[ N^{-1} \sum_i^N \left( f(\theta^{(i)}) - \hat{E}_N[f(\theta)|y] \right)^2 \right]^{1/2}. \quad (53)$$

Another important property of the simulation estimate (52) is that, under the same conditions, the central limit theorem ensures that:

$$N^{1/2} \frac{\left[ \hat{E}_N(f(\theta)|y) - E(f(\theta)|y) \right]}{\left[ \text{var}(f(\theta)|y) \right]^{1/2}} \xrightarrow{d} N(0,1). \quad (54)$$

By means of Monte Carlo integration, it is also possible to marginalise out nuisance parameters and to obtain the relevant marginal posterior pdfs. The simplest way to achieve this is to define  $f(\theta) = p(\theta_1|\theta_2, y)$ . In this way, we have:

$$p(\theta_1|y) = \int p(\theta_1|\theta_2, y) p(\theta_2|y) d\theta_2. \quad (55)$$

The above quantity is estimated for a grid of values of  $\theta_1 \in \Theta_1 \subseteq \mathbf{R}_{k1}$  as:

$$\hat{p}_N(\theta_1|y) = N^{-1} \sum_{i=1}^N p(\theta_1|\theta_2^{(i)}, y). \quad (56)$$

The standard error of the estimate would then be obtained as in expression (53).

An important issue is the efficiency of the simulation estimates. Sometimes it is possible to apply methods that allow one to improve the efficiency of the estimates, without increasing the number of replications being used in the simulation. If the aim of the researcher is to estimate  $E[f(\theta)|y]$ , the most straightforward way is to use the simulated sample mean (52). A more efficient estimate could be obtained by noting that:

$$E(f(\theta)|y) = \int E(f(\theta)|\theta_2, y) p(\theta_2|y) d\theta_2 \quad (57)$$

If the analytical expression of  $E(f(\theta)|\theta_2, y)$  is known, it is possible to estimate  $E(f(\theta)|y)$  in a different way:

$$\hat{E}_T[f(\theta|y)] = N^{-1} \sum_{i=1}^N E[f(\theta|\theta_2^{(i)}, y)] \quad (58)$$

For any given  $N$ , the variance of the estimator (58) is bound to be no larger than the one of the estimator (52); this property holds because of the Rao-Blackwell theorem. In case the analytical expression of  $E(f(\theta)|\theta_2, y)$  is not known, Geweke (1988) describes another method to decrease the standard error of the simulation estimate, called antithetic sampling. If the aim of the researcher is to obtain an estimate for  $E(f(\theta)|y)$ , and the posterior distribution  $p(\theta|y)$  is symmetric around its mode  $\mathbf{v}$ , such that  $p(\theta|y) = p(2\mathbf{v} - \theta|y)$ ,  $E(\theta|y)$  is estimated by antithetically sampling around the mode:

$$\hat{E}_T[f(\theta|y)] = N^{-1} \sum_{i=1}^{N/2} [f(\theta^{(i)}) + f(2\mathbf{v} - \theta^{(i)})] \quad (59)$$

Geweke (1988) shows that the numerical variance of the antithetically sampled estimator is never higher than the variance associated with the simple estimator (52). This is true in the case of symmetric posterior densities, but it is supposed to hold also when the posterior pdf is approximately symmetric.

The Monte Carlo integration principle can also be used in order to estimate posterior probabilities. For instance, suppose that  $h(\theta)$  is defined over  $\Xi \subseteq \mathbb{R}$ ; the probability  $p(h(\theta) \in A|y)$ ,  $A \subseteq \Xi$ , can be evaluated as:

$$E\{I[\theta, h(\theta) \in A] | y\} = p(h(\theta) \in A | y), \quad (60)$$

where  $I[\theta, h(\theta) \in A]$  is the indicator function. Therefore, defining  $f(\theta) = I[\theta, h(\theta) \in A]$ , the sample mean of the indicator functions will give a consistent estimate of the required posterior probability.

Similarly, an aptly defined function  $f(\theta)$  is used to show how one could obtain estimates of the quantiles of the posterior distribution of any function  $h(\theta)$ . All that is required is to sort the draws  $h(\theta^{(i)})$ , and use as estimate of the  $\alpha$  % quantile the value  $h(\theta^{(i)})$  that leaves  $\alpha$  % of the draws on its left. Such an estimate is easily shown to be consistent.

An important aspect to stress is that it is not always very easy to simulate posterior pdfs. Some distributions can be promptly simulated, and this is the case of univariate, and multivariate normal and uniform distributions. Standard distributional results allow easy simulation of those distributions which stem from transformations of the normal distribution. This is the case of univariate and multivariate Student- $t$  distributions, of the  $\chi^2$  distribution, and of the Wishart distribution. Other univariate distributions can be easily simulated by means of direct inversion of the associated cdf: when the cdf  $P(\theta | y)$  has known analytical form, a draw  $u^{(i)}$  on the uniform  $U(0,1)$  distribution can be mapped on a draw  $\theta^{(i)}$  on  $p(\theta | y)$ :

$$\theta^{(i)} = P^{-1}(u^{(i)} | y). \quad (61)$$

As for univariate distributions, when no analytical results are available, it is possible to resort to the method known as "rejection sampling", and described in Devroye (1986), Ripley (1987) and Geweke (1994). Suppose that  $p(\theta | y)$  is the kernel of the posterior pdf the researcher wants to simulate. A comparison function, also called "envelope function" is chosen, with kernel  $g(\theta | \psi)$ , which

depends on a vector  $\psi$  of hyperparameters. The function  $g(\cdot)$  must easily allow simulation. Then, a single draw from  $g(\theta|\psi)$ , say  $\theta^{(i)}$  is retained or rejected on the basis of the outcome of another independent random drawing from the uniform distribution defined over the support :

$$S = \left[ 0, \max_{\theta} \left( \frac{p(\theta|y)}{g(\theta|\psi)} \right) \right] \quad (62)$$

If the result from such drawing, say  $u^{(i)}$ , is less than  $p(\theta^{(i)}|y)/g(\theta^{(i)})$ , then the draw  $\theta^{(i)}$  is accepted, and rejected otherwise. This implies that for any subset  $A$  of the support of  $\theta$ , the probability of getting retained draws is given by:

$$\int_{\theta \in A} g(\theta) \left[ \frac{\frac{p(\theta|y)}{g(\theta)}}{\max_{\theta} \left( \frac{p(\theta|y)}{g(\theta)} \right)} \right] d\theta \propto \int_{\theta \in A} p(\theta|y) d\theta$$

i.e. the algorithm generates synthetic draws from the target distribution  $p(\theta|y)$ , via the comparison function. The problem is how to optimally choose the comparison function. Generally, the aim is to maximise computational efficiency, i.e. to maximise the unconditional probability of retaining draws from the comparison function. We have thus to choose the hyperparameters of the envelope function so as to solve:

$$\min_{\psi} \left[ \max_{\theta} \left( \frac{p(\theta|y)}{g(\theta|\psi)} \right) \right] \quad (63)$$

Notice that the method described does not require knowledge of the normalising constants associated with  $p(\theta | y)$  and  $g(\theta | \psi)$ . In the next chapter a detailed application of this sampling method is fully described.

When the posterior pdf is multivariate with an analytically untractable form, it is possible to resort to *importance sampling*. Importance sampling was discussed by Hammersley and Handscomb (1964, Section 5.4), and its use in econometrics was sparked by the works of Kloek and van Dijk (1978) and by Geweke (1989). The approach is relatively simple and can be summarised as follows. Suppose that the joint posterior pdf  $p(\theta|y)$  is not tractable. An "importance function"  $I(\theta)$  is specified with the following properties. Firstly, it must be possible to obtain draws from it. Secondly, the support of  $I(\theta)$  must include the support of  $p(\theta|y)$ . It is possible to write the posterior expectation of any function  $f(\theta)$  of the parameters as:

$$E[f(\theta)|y] = \int f(\theta) p(\theta|y) d\theta = \int f(\theta) w(\theta) I(\theta) d\theta, \quad (64)$$

$$w(\theta) = p(\theta|y) / I(\theta)$$

Under the conditions stated above, the required posterior expectation can be numerically estimated as follows:

$$\hat{E}_T[f(\theta)|y] = \left[ \sum_{i=1}^N f(\theta^{(i)}) w(\theta^{(i)}) \right] / \left[ \sum_{i=1}^N w(\theta^{(i)}) \right]. \quad (65)$$

In fact, it is possible to show that the simulation estimator (65) converges in probability to  $E[f(\theta)|y]$  (see Geweke, 1989). The problem is that sometimes it could prove necessary to resort to an unreasonably high number of simulations to achieve acceptable precision in the estimate. From this point of view, it is useful to



describe the asymptotic distribution of the numerical estimate (65). In order to do that, Geweke (1989) adds a further condition: the posterior expectations  $E[w(\theta)|y] = \int w(\theta) p(\theta|y) d\theta$  and  $E[(g(\theta))^2 w(\theta)|y] = \int (g(\theta))^2 w(\theta) p(\theta|y) d\theta$  must be finite. Defining the quantity:

$$\hat{\sigma}_N = \left\{ \left[ \sum_{i=1}^N (f(\theta^{(i)}) - \hat{E}_N(f(\theta)|y)) \right]^2 / \left[ \sum_{i=1}^N w(\theta^{(i)}) \right]^2 \right\}^{1/2}, \quad (66)$$

which is termed "numerical standard error, then via central limit theorem it is possible to state (see Geweke, 1989, Theorem 2):

$$N^{1/2} [\hat{E}_N(f(\theta)|y) - E(f(\theta)|y)] \xrightarrow{d} N(0, \sigma^2),$$

$$N\hat{\sigma}_N^2 \rightarrow \sigma^2, \text{ where:} \quad (67)$$

$$\sigma^2 = \frac{\int [f(\theta) - E(f(\theta)|y)]^2 w(\theta) p(\theta|y) d\theta}{\left[ \int p(\theta|y) d\theta \right] \left[ \int I(\theta) d\theta \right]}.$$

In this way it is immediately possible to give a measure of the precision in the simulation estimate of  $E(f(\theta)|y)$ . This measure is called relative numerical efficiency (RNE):

$$RNE = \left\{ N^{-1} \left[ \sum_{i=1}^N (f(\theta^{(i)}) - \hat{E}_N(f(\theta)|y))^2 \right] \right\} / (N\hat{\sigma}_N^2). \quad (68)$$

The numerator of this ratio is the variance of the estimate of  $E(f(\theta)|y)$  that would result were  $p(\theta|y)$  directly simulable, i.e. if  $I(\theta) = p(\theta|y)$ . The denominator is the estimate of the variance resulting from the use of an importance function different

from  $p(\theta|y)$ . Low values of the ratio indicate a poor choice of the importance function.

In summary, in order to obtain precise results it is necessary to choose  $I(\theta)$  in a very accurate way; this choice calls for a very thorough understanding of the main features of the joint posterior distribution  $p(\theta|y)$ . As a necessary condition to obtain reliable simulation estimates, the importance function must have fatter tails than the joint posterior pdf, otherwise values of  $\theta$  in the tails of  $I(\theta)$  would be associated with extremely large values of  $w(\theta)$ . This would attach very high weights to the corresponding values of  $f(\theta)$ . In this way, the resulting estimate could be dominated by infrequent draws on  $w(\theta)$ . Sometimes, when the features of the joint posterior pdf  $p(\theta|y)$  are completely unknown, the choice of  $I(\theta)$  can only be tentative, and no feasible solution could be acceptable.

For all these reasons, in recent years other simulation techniques have become increasingly popular. These techniques, known as Markov Chain Monte Carlo methods (*MCMC*) permit Monte Carlo integration without resorting to the importance sampling principle, allowing draws to be obtained from joint posterior distributions that are otherwise intractable. These *MCMC* methods are reviewed in the next sub-section.

#### **[3.4.d] Markov Chain Monte Carlo: Gibbs Sampling and Metropolis-Hastings Algorithms.**

Suppose that it is required to draw from the density  $p(\theta)$ , where  $\theta$  is a  $n$ -dimensional random variable with support  $\Omega \subseteq \mathbb{R}^n$ . If  $p(\theta)$  does not allow *i.i.d.* drawing, it is possible to use dependent samples generated by a Markov chain having  $p(\cdot)$  as equilibrium distribution. Exhaustive and up-to-date surveys on this issue are Tierney, (1991, 1994) and Chib and Greenberg (1994 a).

A time-homogeneous Markov chain (henceforth *MC*) is defined as a sequence of random variables  $\{\theta_i, i=1, 2, \dots, N\}$  such that:

$$p(\theta_{i+1}|\theta_0, \theta_1, \dots, \theta_i) = p(\theta_{i+1}|\theta_i) = p(\theta_{j+1}|\theta_j), i, j = 1, 2, \dots, N, \quad (69)$$

where  $p(\theta_{i+1}|\theta_i)$  is the conditional density with respect to  $\mu(\cdot)$ , a  $\sigma$ -finite measure on the Borel  $\sigma$ -field generated on  $\Omega$ . The Markov property is expressed by the first of the two equalities above. The conditional density  $p(\theta_{j+1}|\theta_j)$  will be referred to as  $p(\theta_j, \theta_{j+1})$ .

We define  $\mathbf{x} = \theta_i$ ,  $\mathbf{y} = \theta_{i+1}$ , and we have that the probability of  $\mathbf{y}$  belonging to  $A \subseteq \Omega$  conditional on  $\mathbf{x}$  is <sup>2</sup>:

$$P(\mathbf{x}, d\mathbf{y}) = p(\mathbf{x}, \mathbf{y})\mu(d\mathbf{y}), d\mathbf{y} = \{\mathbf{y}: \mathbf{y} \in A\}. \quad (70)$$

The expression above defines the (one step ahead) *transition kernel* of the *MC*, and similarly it is possible to define its  $n$ -step ahead counterpart :

$$P^n(\mathbf{x}, d\mathbf{y}) = \text{prob}(\theta_{i+n} \in A | \mathbf{x}), n \geq 1. \quad (71)$$

The *invariant* distribution of *MC* is defined as:

$$\pi^*(d\mathbf{y}) = \pi(\mathbf{y})\mu(d\mathbf{y}) = \int \pi(\mathbf{x})P(\mathbf{x}, d\mathbf{y}) \mu(d\mathbf{x}), \quad (72)$$

---

<sup>2</sup> In this analysis, I do not assign positive probability mass to any  $\mathbf{x}$ . Generalising the context above to cover such cases would yield:

$$P(\mathbf{x}, d\mathbf{y}) = p(\mathbf{x}, \mathbf{y}) \mu(d\mathbf{y}) + \delta(\mathbf{x}) [1 - p(\mathbf{x}, \mathbf{y}) \mu(d\mathbf{y})],$$

where  $p(\mathbf{x}, \mathbf{x})=0$  and  $\delta(\mathbf{x})$  is the point mass attributed to  $\mathbf{x}$ .

and can be interpreted as the unconditional distribution of  $y$ . Such a distribution is shown to exist (see Tierney, 1994, Section 3) if the MC is *reversible*, i.e. if :

$$\pi(\mathbf{x}) p(\mathbf{x}, y) = \pi(y) p(y, \mathbf{x}), \quad (73)$$

which has to be interpreted as requiring 'well behaviour' of the joint density.

At this stage, the invariant distribution  $\pi^*(dy)$  is the *equilibrium* distribution of the MC if:

$$\lim_{n \rightarrow \infty} P^n(\mathbf{x}, dy) = \pi^*(dy), \quad (74)$$

for every measurable set  $dy$  and for any  $\mathbf{x}$ . The condition above requires that, starting off the MC at any point  $\mathbf{x}$ , the  $n$ -step ahead conditional distribution converges to the invariant distribution  $\pi^*(y)$ .

Two important properties of a MC are defined as follows:

a) *Irreducibility*: a MC is said to be  $\pi^*$ -irreducible if  $\forall \mathbf{x} \in A$  such that  $\pi^*(A) > 0$ , then we have  $\text{prob}(\theta_i \in A | \theta_0 = \mathbf{x}) > 0$  for some  $i > 0$ . This property means that starting from any state having positive probability according to the invariant distribution, the chain can return there with positive conditional probability.

b) *Aperiodicity*: a MC is said to be *aperiodic* if no partition of  $\Omega$  defined as  $\{\Omega_0, \Omega_1, \dots, \Omega_p, p \geq 2\}$  exists, such that:  $P(\theta_{i(p)} \in \Omega_k | \theta_0 \in \Omega_0) = 1$ , with  $k=1, \dots, p-1$ . This condition amounts to ruling out that the MC might deterministically visit subsets of the support of  $\pi^*$  at regular intervals.

Tierney (1994, Section 3) has proved the following *ergodicity* result, stating that, if the MC defined by  $P(\mathbf{x}, dy)$  has invariant distribution  $\pi^*(dy)$  and is irreducible, then  $\pi^*(dy)$  is unique. Moreover if the MC is also aperiodic, then:

- a)  $\pi^*(dy)$  is the equilibrium distribution of the MC;
- b) for any function  $h(\theta)$ , we have:

$$N^{-1} \sum_{i=1}^N h(\theta^{(i)}) \Rightarrow \int h(\mathbf{x}) \pi^*(\mathbf{x}). \quad (75)$$

These results mean that aperiodicity and irreducibility of the chain ensure convergence to the invariant distribution, and that the sample of dependent draws of  $h(\cdot)$  taken from the *MC* converges to the theoretical expected value taken with respect to the invariant density. This is a generalisation of the law of large numbers applied to *MC*'s.

Drawing from an intractable target density  $\pi(\theta)$  can therefore be achieved by aptly defining a *MC* drawing scheme that has  $\pi(\theta)$  as equilibrium distribution. A very convenient way to define such a *MC* scheme is to resort to conditioning the target density. This approach, known as Gibbs Sampling (*GS*), is being increasingly applied in the Bayesian literature since it is conceptually very simple and very easy to implement. Geman and Geman (1984) introduced the technique, and Gelfand and Smith (1990) and Smith and Roberts (1993) give interesting discussions about the interrelation between different numerical methods. Chib and Greenberg (1994b) give a wide range of econometric application of *GS*.

The idea behind *GS* is quite simple and intuitive. Suppose that all we know is the mathematical expression for an analytically untractable density. Suppose further that the conditional posterior distributions of an exhaustive collection of  $k$  mutually exclusive subsets of the parameter vector:  $\theta = [\theta_1', \theta_2', \dots, \theta_k']'$  are "available" in the sense that each of them can be easily simulated.

*GS* works as follows. We start from an arbitrary initialisation of the parameter vector:

$$\theta^{(0)} = [\theta_1^{(0)'}, \theta_2^{(0)'}, \dots, \theta_k^{(0)'}]'. \quad (76)$$

Each pass of the algorithm consists of  $k$  steps. At the  $i^{\text{th}}$  step of the first pass,  $i=1, 2, \dots, N$ , a random draw is obtained from the conditional density:

$$p(\theta_i | \theta_1^{(i-1)} \dots \theta_{i-1}^{(i-1)}, \theta_{i+1}^{(i)} \dots \theta_k^{(i)}), i = 1, 2, \dots, k, \quad (77)$$

Clearly, each pass of *GS* represents a step in a *MC*. Calling  $\mathbf{x} = \theta_i$ ,  $\mathbf{y} = \theta_{i+1}$ , the transition kernel has density:

$$p(\mathbf{x}, \mathbf{y}) = \prod_{j=1}^k p(\mathbf{y}_j | \mathbf{y}_1, \dots, \mathbf{y}_{j-1}, \mathbf{x}_{j+1}, \dots, \mathbf{x}_k). \quad (78)$$

Applying Bayes' theorem to each factor above yields:

$$p(\mathbf{y}_j | \mathbf{y}_1, \dots, \mathbf{y}_{j-1}, \mathbf{x}_{j+1}, \dots, \mathbf{x}_k) = \frac{p(\mathbf{y}_j | \mathbf{y}_1, \dots, \mathbf{y}_{j-1}) p(\mathbf{x}_{j+1}, \dots, \mathbf{x}_k | \mathbf{y}_1, \dots, \mathbf{y}_j)}{p(\mathbf{x}_{j+1}, \dots, \mathbf{x}_k | \mathbf{y}_1, \dots, \mathbf{y}_{j-1})}, \quad (79)$$

Choosing  $\mu(\cdot)$  as the Lebesgue measure and plugging (79) into (72) gives:

$$\int \pi(\mathbf{x}) P(\mathbf{x}, d\mathbf{y}) d\mathbf{x} = \pi(\mathbf{y}) d\mathbf{y} \int \prod_{j=1}^k p(\mathbf{x}_j | \mathbf{y}_1, \dots, \mathbf{y}_{j-1}, \mathbf{x}_{j+1}, \dots, \mathbf{x}_k) d\mathbf{x} = \pi^*(\mathbf{y}), \quad (80)$$

showing that *GS* has the target distribution as invariant distribution.

In order to show that the equilibrium distribution is indeed  $\pi^*(\mathbf{y})$ , aperiodicity and irreducibility of the Gibbs sampling *MC* should be proved. In order to ensure that these properties hold, different sets of sufficient conditions have been proposed (Chan, 1993, Roberts and Smith, 1994, Tierney, 1991, 1994). As for *GS*, Tierney put forward the following conditions on the conditional distributions:

$$(1) p(\theta_i | \theta_{\cdot}, i \neq j) > 0 \quad \forall \theta_i \in \Omega_i, \theta_j \in \Omega_j,$$

$$(2) P(\theta_j \in A \mid \theta_i, i \neq j) > 0 \quad \forall P_j\text{-measurable set } A \in \Omega_j$$

$$(3) P(\theta_j \in A \mid \theta_i, i \neq j) \text{ is a continuous function of } \theta_i, i \neq j.$$

These conditions are easily shown to hold in most of the econometric applications, in this way ensuring that the *MC* defined by use of *GS* converges to the target density.

Some complications might arise when one or more conditional distributions do not allow random drawing. In this case it is possible to conveniently use another *MC* based sampling method, known as "Metropolis-Hastings" (*MH*) algorithm. For a description of the algorithm, see Tierney (1991, 1994), Chib and Greenberg (1994a). A *MH* algorithm to simulate  $\pi(\theta)$  works as follows: starting from  $\mathbf{x} = \theta^{(i)}$ , a draw  $\mathbf{y} = \theta^{(i+1)}$  is generated from a 'candidate' density  $q(\mathbf{x}, \mathbf{y})$ . The transition kernel is then:

$$Q(\mathbf{x}, d\mathbf{y}) = q(\mathbf{x}, \mathbf{y}) \mu(d\mathbf{y}). \quad (81)$$

This draw is either retained, or discarded (setting  $\mathbf{y} = \mathbf{x}$  in the following step of the algorithm), subject to a further dichotomic randomisation with probability:

$$\begin{aligned} \alpha(\mathbf{x}, \mathbf{y}) &= \frac{\pi(\mathbf{y})q(\mathbf{y}, \mathbf{x})}{\pi(\mathbf{x})q(\mathbf{x}, \mathbf{y})} \quad \text{if } \pi(\mathbf{x})q(\mathbf{x}, \mathbf{y}) > 0, \\ &= 1 \quad \text{if } \pi(\mathbf{x})q(\mathbf{x}, \mathbf{y}) = 0. \end{aligned} \quad (82)$$

Indicating with  $\mu(d\mathbf{y})$  the Lebesgue measure and defining the transition kernel as:

$$P(\mathbf{x}, d\mathbf{y}) = q(\mathbf{x}, \mathbf{y}) \alpha(\mathbf{x}, \mathbf{y}) \mu(d\mathbf{y}), \quad (83)$$

expression (72) for the *MH* chain becomes:

$$\alpha(\mathbf{x}, \mathbf{y}) = \int \pi(\mathbf{x})q(\mathbf{x}, \mathbf{y}) \frac{\pi(\mathbf{y})q(\mathbf{y}, \mathbf{x})}{\pi(\mathbf{x})q(\mathbf{x}, \mathbf{y})} d\mathbf{y}d\mathbf{x} = \pi^*(d\mathbf{y}), \quad (84)$$

showing that the invariant distribution of the *MH* Markov chain is  $\pi^*(\mathbf{y})$ . The sufficient conditions for the *MH* chain to converge to  $\pi^*(\mathbf{y})$  (see Mengersen and Tweedie, 1993) are very mild and easy to verify: if  $\pi(\mathbf{x})$  and  $q(\mathbf{x}, \mathbf{y})$  are positive and continuous  $\forall \mathbf{x}, \mathbf{y} \in \Omega$ , then the chain is irreducible and aperiodic and the Tierney's ergodicity results hold.

The main issue in the implementation of the *MH* algorithm is connected to the choice of  $q(\mathbf{x}, \mathbf{y})$ . As detailed in Tierney, (1994, Section 2.3), different specifications of  $q(\mathbf{x}, \mathbf{y})$  lead to variants of the basic *MH*. In practice, it is necessary to choose carefully the candidate function, in order to run the *MH* sampler efficiently, otherwise the acceptance rate of the draws from  $q(\cdot)$  could prove to be very low.

In some cases, it could be profitable to resort to a mixed *MH-GS* sampling scheme: for instance consider the case when  $p(\theta_i | \theta_{-i}, i \neq j)$  cannot be simulated, while the other conditional densities can. Then, at each step of the *GS* involving  $p(\theta_i | \theta_{-i}, i \neq j)$ , the required draw can be obtained by means of an aptly defined *MH* algorithm.

In synthesis, having designed a *MC* drawing scheme having  $\pi(\theta)$  (the required posterior pdf) as its invariant distribution, and having checked that the conditions ensuring convergence to  $p(\theta | \mathbf{y})$  are fulfilled, a string of  $N$  draws  $\theta^{(i)}$  can be obtained, after discarding first an initial batch of  $N_0$  passes, in order to allow for convergence to occur. The draws being retained obtained in the  $N$  subsequent passes of *GS* can be used to apply the Monte Carlo integration principle in order to estimate by simulation the posterior moments of any  $f(\theta)$ . The estimate is obtained as the sample mean of the  $f(\theta^{(i)})$ ,  $i = N_0+1, \dots, N_0+N$ , as in expression (52).

In the next sub-section, I describe how to measure the accuracy of the resulting *MC* Monte Carlo estimates.



### [3.4.e] Measuring the accuracy of MC Monte Carlo estimates

The most natural measure of the accuracy of  $\bar{f}_T = \hat{E}_T[f(\theta)|y]$ , the simulation estimate of the posterior expectation of  $f(\theta)$ , is clearly its Monte Carlo standard error. If the random draws on  $f(\theta)$  were *i.i.d.*, we could associate an easy estimate of this standard error, namely:

$$\hat{\sigma}_{i.i.d.}(\bar{f}_T) = \left\{ N^{-1} \sum_{j=1}^N [f(\theta^{(j)}) - \bar{f}_T]^2 / (N-1) \right\}^{1/2} \quad (85)$$

Unfortunately, MC draws on  $f(\theta)$  are inherently autocorrelated. This context is very similar to regression analysis with non-spherical errors, where the first order moment estimates are consistent, but the second order moment estimates are not. It is therefore profitable to resort to the same kind of heteroskedasticity-autocorrelation consistent (HAC) estimators of the standard error of the sample mean of  $f(\theta)$  (see Newey and West, 1987, Andrews, 1991, and Andrews and Monahan, 1992). The main idea behind this class of estimators is the central limit theorem for dependent processes, stating that:

$$N^{1/2} \{ \bar{f}_T - E[f(\theta)|y] \} \xrightarrow{d} N(0, S_f(0)), \quad (86)$$

where  $S_f(0)$  is  $2\pi$  times the spectral density function of  $f(\theta)$  at frequency zero. This property holds also in the case of the Monte Carlo estimator based on a MC sample. Therefore, a HAC estimator of the standard error of  $\hat{E}_T[f(\theta)|y]$  is the one based on a consistent estimate of its spectral density function at frequency zero. The literature has put forward different choices in this respect. The simplest one, which delivers a well behaved estimate of the standard error, is the Newey and

West estimator, already mentioned in Chapter 2, based on a time domain estimator of  $S_{\lambda}(0)$ , with a fixed bandwidth and a Bartlett window:

$$\begin{aligned}\hat{\sigma}_{NW}^2(\bar{f}_T) &= N^{-1} \sum_{j=-m}^m w(j, m) \tilde{\gamma}_j(f(\theta)), \\ w(j, m) &= \max[0, 1 - |j| / (m + 1)], \\ \hat{\gamma}_j(f(\theta)) &= N^{-1} \sum_{i=j+1}^N w(j, m) [f(\theta^{(i)}) - \bar{f}_T] [f(\theta^{(i-j)}) - \bar{f}_T].\end{aligned}\tag{87}$$

Following the same argument as with the importance sampling Monte Carlo evaluations, a *RNE* index can be constructed as:

$$RNE = \hat{\sigma}_{iid}(\bar{f}_T) / \hat{\sigma}_{NW}(\bar{f}_T).\tag{88}$$

High values of this index indicate good reliability of the Monte Carlo estimate based on the Gibbs sample. In fact, when *RNE* is high, it turns out that the Gibbs sample draws on  $f(\theta)$  are not very much correlated, and therefore a precise estimation of  $E(f(\theta)|y)$  can be easily obtained. If *RNE* is low, this means that the desired level of precision can be obtained only by increasing the number of draws by a factor of  $(RNE)^{-1}$ .

Another important problem is that of assessing whether or not the *MC* has converged to the desired joint posterior pdf in a certain finite sample. The issue is in fact very problematic and in the literature different procedures have been proposed. A sensible way of proceeding is to observe the sampling output of the *MC* and from it infer whether convergence has occurred or not. In this respect, Ritter and Tanner (1992) propose to evaluate at each step the ratio between the target density and its estimate computed at each pass of the *MC*; stability of this ratio indicates convergence has been reached. Gelman and Rubin (1992) propose to run different chains starting from different initial states, and comparing the

sample variability within and across the single chains. Geweke (1992) constructs a diagnostic test based on the premise that once convergence has occurred, the subsequent draws will have all the same distribution. Therefore, having discarded a batch of preliminary draws, in order to "warm up" the algorithm, it is possible to check convergence of the chain to the posterior distribution of any function  $f(\theta)$  by means of:

$$CD(f(\theta)) = [\bar{f}_A(\theta) - \bar{f}_B(\theta)] / [\tilde{\sigma}_{NW}(\bar{f}_A(\theta)) + \tilde{\sigma}_{NW}(\bar{f}_B(\theta))]^{1/2} \quad (89)$$

This is a *HAC* test of equality between the sample mean of the early and the late draws in the sampling scheme ( $\bar{f}_A(\theta)$  and  $\bar{f}_B(\theta)$  respectively), sufficiently far apart to neglect the covariance between the two. The null hypothesis of equality of the means of the two sample periods can be tested by using the asymptotic normality of (89).

### [3.5] Bayesian Analysis of Non Stationary Univariate Models

This section describes how non-stationary univariate linear time series models can be treated in a Bayesian framework. I start from the simple, no-deterministics *AR*(1) model:

$$y_t = \rho y_{t-1} + e_t, \quad e_t \sim N.i.d.(0, \sigma^2). \quad (90)$$

The unknown parameters in the model are  $\rho$  and  $\sigma^2$ , and the likelihood function of a sample of  $T$  observations is very simple, especially when we condition on the first observation  $y_0$ :

$$p(y|y_0, \rho, \sigma^2) = (2\pi\sigma^2)^{-T/2} \exp\{-1/(2\sigma^2)(e'e)\}, \quad e = y - \rho y_{-1}. \quad (91)$$

It is easy to see that a straightforward natural conjugate analysis is possible, given that sufficient statistics for  $\rho$  and  $\sigma^2$  do exist. By specifying a normal-inverted Gamma prior for the unknown parameters, as in expressions (35) and (36), one obtains a marginal Student- $t$  distribution for  $\rho$ :

$$p(\rho|y) \propto \left[ \nu^* + (\rho - \bar{\rho})^2 / \bar{\sigma}^2 \right]^{-(\nu^*+1)/2}$$

$$\nu^* = T + \nu_0, \quad \bar{\rho} = q^* [y_{-1}' y_{-1} + c_0^{-1} \rho_0], \quad q^* = [y_{-1}' y_{-1} + c_0^{-1}]^{-1},$$

$$\bar{\sigma}^2 = \{ \nu_0 s_0^2 + y' y + \rho_0^2 c_0^{-1} + \bar{\rho}^2 / q^* \} / (\nu^*)$$
(92)

What we get is therefore a symmetric, analytically tractable marginal posterior pdf. Higher dimensional AR( $p$ ) processes with deterministic components with the simple linear parameterisation of the kind:

$$\rho(L) y_t = \alpha + \beta t + e_t$$
(93)

can be analysed in exactly the same way. A natural conjugate prior exists, since we have sufficient statistics for all the parameters of the model. The posterior pdf for the autoregressive parameters is  $p$ -variate Student- $t$ , as we have seen for the linear regression model.

Things get a bit more involved when dealing with processes with a moving average part. We take as an example the simple MA(1) process:

$$y_t = e_t + \theta e_{t-1}$$
(94)

The likelihood function for a sample of dimension  $T$  is somewhat more complicated than for an  $AR(p)$  process. Given the autocorrelation structure of the process, the likelihood function is:

$$p(\mathbf{y}|\theta, \sigma) = (2\pi\sigma^2)^{-T/2} |\mathbf{B}|^{-1/2} \exp\{-1/(2\sigma^2)\mathbf{y}'\mathbf{B}^{-1}\mathbf{y}\}, \quad (95)$$

$$\mathbf{B}_{(T \times T)} = \begin{bmatrix} 1+\theta^2 & \theta & \dots & 0 & 0 \\ \theta & 1+\theta^2 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1+\theta^2 & \theta \\ 0 & 0 & \dots & \theta & 1+\theta^2 \end{bmatrix}$$

Sufficient statistics for  $\theta$  cannot be obtained and therefore no conjugate prior pdf exists. Whichever prior pdf is implemented, the resulting posterior pdf needs to be treated with numerical simulation techniques (see Chib and Greenberg, 1994 *b*).

The number of problems encountered in the frequentist analysis of non-stationary time series described in Chapter 2 induced many researchers to use Bayesian techniques in order to discriminate between competing models of non-stationarity. In Sims (1988), this framework is adopted for the first time, in order to decide whether an  $AR(1)$  process has a unit root or not.  $H_0$  is taken to be  $\rho = 1$ , while  $H_1$  embeds stationarity. Therefore a sharp null is compared to a composite alternative. In order to reflect neutrality of the researcher, Sims proposes a unit prior odds ratio. The prior suggested under the alternative is flat over a subset of the stationarity region. Its width is determined on the basis of the observation frequency of the data under study: the higher the frequency, the more concentrated near one is the prior.

In Schotman and Van Dijk (1991*a* and *b*) extensions and applications of the Sims approach are provided. The simplest model which is considered therein is again an

$AR(1)$  process without deterministic component. The two hypotheses are specified as in Sims' analysis:

$$H_0: \rho = 1, \text{ with } p(H_0) = p,$$

$$H_1: \rho \in S, \text{ with } p(H_1) = 1-p, S = \{\rho, -1 < \alpha < \rho < 1\}.$$

Under  $H_0$ ,  $\rho$  is equal to 1 with probability equal to one, and under  $H_1$  a flat prior is specified for  $\rho$  defined over the range given by  $S$ :

$$p(\rho|H_1) = (1-\alpha)^{-1}. \quad (96)$$

Therefore if the random variable 'true hypothesis' is marginalized out, a peculiar distribution for  $\rho$  is obtained: it is uniform and continuous in  $S$  and it assigns a positive probability mass to the event  $\rho = 1$ . This is necessary, given that the aim is to compare a sharp point to a composite hypothesis. The other unknown parameter in the model,  $\sigma$ , is treated as a nuisance parameter and it is provided with the usual Jeffreys prior:

$$p(\sigma|H_0) = p(\sigma|H_1) \propto \sigma^{-1}. \quad (97)$$

The posterior odds ratio is obtained by means of analytical integration, yielding:

$$k_1 = k_0 \frac{1-\alpha}{s_{\hat{\rho}}} \left[ \frac{\sigma_0}{\hat{\sigma}} \right] \frac{B[(T-1)/2, 1/2]}{(T-1)^{1/2}} \left[ F_{T-1} \left( \frac{(1-\hat{\rho})}{s_{\hat{\rho}}} \right) - F_{T-1} \left( \frac{(\alpha-\hat{\rho})}{s_{\hat{\rho}}} \right) \right]^{-1}, \quad (98)$$

where  $\sigma_0$  is the standard error of the differenced series,  $\hat{\sigma}$  is the standard error of the  $H_1$  autoregression,  $\hat{\rho}$  is OLS estimate of  $\rho$  and  $s_{\hat{\rho}}$  the associated standard error,  $B(\cdot, \cdot)$  is the Beta function, and  $F_{T-1}(z)$  is the  $T-1$  degree of freedom Student  $t$  cdf evaluated at  $z$ .

In their prior specification, Schotman and van Dijk partly follow Sims, by specifying a balanced prior odds ratio. On the other hand, the lower bound of the

stationarity support of  $\rho$ ,  $\alpha$ , is chosen in a different way, in order to accomplish the important aim of balancing two risks. The first is to define too small a stationary interval, in this way failing to consider values of  $\rho$  with non-negligible likelihood; the second is to choose it too large and induce the denominator of  $k_1$  to get very low and to favour the null independently of any data evidence. As Schotman and van Dijk point out, this phenomenon can be given a 'Lindley paradox' (see Lindley, 1957) interpretation: as the sample size increases, Bayesian analysis fails to yield results converging to the ones obtained by resorting to sampling theory techniques. The choice of  $\alpha$  is then made on the basis of the following considerations: given that under  $H_1$  the marginal posterior pdf of  $\rho$  is Student-t with mean  $\hat{\rho}$  and scale factor  $s_\rho$ ,  $\alpha$  is identified by the requirement that it cover 99% of such a distribution truncated at  $\rho=1$ .

An important point to note is that, given the way in which the prior distributions are defined under both hypotheses, there is a smooth continuous transition from  $H_0$  to  $H_1$  in terms of the associated posterior distributions. This means that from  $p(\rho, \sigma|H_1, y)$ , we obtain  $p(\sigma|H_0, y)$ , simply by substituting  $\rho = 1$ . This smooth transition is appealing when compared with the radical asymmetry between the asymptotic distributions of the OLS estimate of  $\rho$  when the model is stationary and when it is not. This seems to be a very valuable result, in the sense that the occurrence of a unit root in the autoregressive representation does not affect dramatically the properties of the finite sample posterior pdf of the remaining parameters in the model.

Schotman and van Dijk (1991a, section 3) consider an augmented model with an intercept term, choosing the non-linear parameterisation. This is done for the reasons of interpretation already discussed:

$$y_t - \mu = \rho (y_{t-1} - \mu) + e_t \quad (99)$$

The same prior pdf's for  $\sigma$  and  $\rho$  are specified under  $H_0$  and  $H_1$ . A problem arises in the treatment of the parameter  $\mu$ , since this parameter is identified only under  $H_1$ . In fact, the adoption of a diffuse prior on  $\mu$  would induce the resulting posterior odds ratio  $k_1$  to go to infinity, independently of the data information. Intuitively, the denominator of the posterior odds ratio goes to zero because values of  $\mu$  associated with small likelihood values are averaged together with equal weights. This does not happen in the numerator, since  $\mu$  does not enter the likelihood function under  $H_0$ . A proper prior distribution is required. Since  $\mu$  is the unconditional mean of the process under  $H_1$ , the more persistent the process, the less accurate will be the data information about  $\mu$ . On the basis of this last consideration, Schotman and Van Dijk specify the following prior distribution for  $\mu$  under  $H_1$ :

$$p(\mu | \rho, H_1) \propto (1-\rho^2)^{1/2} \sigma^{-1} \exp\{-(1-\rho^2)(\mu-y_0)^2/(2\sigma^2)\}. \quad (100)$$

Two things about this prior distribution should be noted. First, the prior mean of  $\mu$  is set equal to the initial observation of the series, and the prior variance is the unconditional variance of a single observation on a  $AR(1)$  process. For this reason when combined with the likelihood function, this prior distribution produces a joint posterior distribution which coincides with the unconditional likelihood of all the observations from  $y_0$  to  $y_T$ . Secondly the prior variance goes to infinity as  $\rho$  goes to one but more slowly than  $1/(1-\rho)^2$ . This is important to ensure the smooth transition from  $H_1$  to  $H_0$ , as will be shown shortly. Since the prior is not natural conjugate, the resulting posterior pdf is non-standard.

Under  $H_0$ , the prior is specified as follows. A Jeffreys prior is assumed for  $\sigma$ . For  $\mu$  a proper arbitrary prior pdf  $g(\mu)$  is specified, but since this parameter does not appear in the likelihood function, it can be thought of as being marginalised out at the outset. Thus, we have:



$$p(H_0, \mu, \sigma | y) \propto p \sigma^{T-1} \exp[-\Delta y' \Delta y / (2\sigma^2)] \quad (101)$$

Integrating out  $\mu$  and  $\sigma$  gives:

$$p(H_0 | y) \propto p \Gamma(T/2) [\Delta y' \Delta y]^{-T/2} \quad (102)$$

Under  $H_1$ , the joint posterior pdf is:

$$p(H_1, \mu, \rho, \sigma | y) \propto \frac{(1-p)}{1-\alpha} \sigma^{-T-2} (1-\rho^2)^{1/2} \exp \left\{ -\frac{1}{2\sigma^2} [\mathbf{e}' \mathbf{e} + (1-\rho^2)(\mu - y_0)] \right\}, \quad (103)$$

$$\mathbf{e} = \mathbf{y} - \rho \mathbf{y}_{-1} - \mu (1-\rho) \mathbf{i},$$

where  $\mathbf{i}$  is a  $(T \times 1)$  vector of ones. By means of analytical integration, the posterior probability of  $H_1$  is readily obtained:

$$p(H_1 | y) \propto \frac{(1-p)}{1-\alpha} \Gamma(T/2) \int_a^1 h(\rho) [S^2(\rho)]^{-T/2} d\rho, \quad (104)$$

$$h(\rho) = [1 + T(1-\rho)/(1+\rho)]^{-1/2},$$

$$S^2(\rho) = [\mathbf{y} - \rho \mathbf{y}_{-1} - \bar{\mu}(\rho)(1-\rho)\mathbf{i}]' [\mathbf{y} - \rho \mathbf{y}_{-1} - \bar{\mu}(\rho)(1-\rho)\mathbf{i}],$$

$$\bar{\mu}(\rho) = [\mathbf{i}'(\mathbf{y} - \rho \mathbf{y}_{-1}) + (1+\rho)y_0] / [(1+\rho)T + (1+\rho)].$$

Note that the integral in (104) does not have any analytical solution, since  $S^2(\rho)$  is not quadratic in  $\rho$ . The posterior odds ratio is given by:

$$k_1 = k_0 \frac{(1-\alpha) [\Delta y' \Delta y]^{-T/2}}{\int_a^1 h(\rho) [S^2(\rho)]^{-T/2} d\rho} \quad (105)$$

In this case too, the denominator of  $k_1$  has a continuous smooth transition to the numerator, when  $\rho$  approaches 1, since  $h(1)=1$ ,  $\bar{\mu}(1) = (y_T+y_0)/2$ , and  $S^2(1)=\Delta y'\Delta y$ . The lower bound  $\alpha$  is determined as before on the basis of the marginal posterior pdf of  $\rho$  under  $H_1$ .

Clearly, model (99) falls short of providing a general dynamic specification and of providing a trend stationary alternative. In Schotman and van Dijk (1991b), a general  $AR(p)$  is specified for the detrended data:

$$\rho(L)(y_t - \mu - \delta t) = e_t, \quad \rho(L) = 1 - \rho_1 L - \dots - \rho_p L^p. \quad (106)$$

The model is then reparameterised as in the *ADF* analysis:

$$\rho^*(L)(y_t - \delta) = -\rho(y_{t-1} - \mu - \delta(t-1)) + e_t, \quad \rho(L) = 1 - \phi_1 L - \dots - \phi_p L^p, \quad (107)$$

where  $\rho(L) = \rho^*(L) \Delta + \rho(1)L$ ,  $\rho = \rho(1)$ . In this context, the unit root hypothesis corresponds to  $\rho = 0$ . In this case, the resulting model is entirely in differences,  $\mu$  vanishes, and a non-zero  $\delta$  gives the drift. No parameter is irrelevant under either hypothesis.

The alternative is formulated as  $H_1: 0 < \rho < \alpha$ , with  $\alpha$  close to zero. As in the previous case, the attention is restricted to an arbitrarily determined sub-set of the range of values the parameter of interest could theoretically assume. Under  $H_1$ , a diffuse prior is specified for all the parameters in  $\rho^*(L)$ , for  $\delta$  and  $\log \sigma$  over the range  $(-\infty, +\infty)$ , while a proper prior has to be specified for  $\mu$ , for reasons discussed above. The prior chosen by Schotman and van Dijk for  $\rho$  is flat and proper over  $(0, \alpha)$ .

Also in this context, since the prior pdf is not natural conjugate, the joint posterior pdf's under  $H_0$  and  $H_1$  call for numerical integration. As for the latter, one can marginalise out analytically all the parameters but  $\rho$  and  $\delta$ , whereas under  $H_0$ , the

posterior pdf calls for numerical integration with respect to  $\delta$ . Schotman and van Dijk (1991b) show that there exists a continuous smooth transition from  $H_1$  to  $H_0$ . This depends on the way in which the prior pdf for  $\mu$  has been specified: the prior variance goes to infinity as  $\rho$  goes to zero but more slowly than  $1/\phi^2$ .

Geweke (1994) extends the analysis of Schotman and van Dijk by using the same parameterisation, but allowing for leptokurtic disturbances. His analysis makes extensive use of *MC* Monte Carlo integration.

An important feature of Bayesian analysis is the possibility of focussing on the posterior distribution of any function of the parameters. A practical example of the relevance of this possibility is immediately provided by following Beveridge and Nelson, (1981): any  $I(1)$  process can be seen as the sum of a temporary and a permanent component. As we have seen in Chapter 2, the permanent component is the cumulated impulse response function for  $\Delta y_t$ :

$$\Delta y_t = \sum_{i=0}^{\infty} c_i e_{t-i} = c(1)e_t \quad (108)$$

The cumulated impulse response is a non linear function of the *AR* parameters, and a linear function of the *MA* parameters. When  $\Delta y_t$  is a pure  $AR(p)$  process, the persistence measure  $c(1)$  has a lower bound given by  $2^{-p}$ . If a *MA* part is introduced in the model, then the lower limit of  $c(1)$  becomes zero. On the basis of these considerations, another possible way to check whether a series is stationary or not could be to construct an *ARMA* representation for its differences, obtain the posterior pdf for  $c(1)$  and see whether the value of zero falls in the *HPD* confidence interval.

### [3.6] Ignorance Priors in Time Series Models

As already pointed out in Sims (1988) and in Sims and Uhlig (1991), the Bayesian approach seems to provide a sensible way to overcome the statistical difficulties implied by the presence of a unit root. These statistical difficulties can be regarded from different points of view: the distributional asymmetry between integrated and non-integrated time series; the observation of confidence sets with a 'disconcerting topology' (Sims (1988), *p.* 466): confidence sets can be disconnected due to the distributional asymmetry between stationarity and integration. As already mentioned, the Bayesian approach, focussing on the posterior distribution of the parameters of interest, gives different solutions. If, for example, in a  $AR(1)$  model,  $\rho$  is provided with a flat prior, the resulting marginal posterior distribution will be Student- $t$ ; the unimodality of this distribution will yield non-disjoint *HPD* confidence sets.

These considerations have led to the adoption of flat priors, not only in the posterior odds ratio testing framework, but also in simulation and applied studies intended to measure the posterior probability of  $\rho$  lying in the stationarity interval. See Sims and Uhlig (1991), DeJong and Whiteman (1989) and (1991). Phillips (1991a) fiercely criticises the use of flat prior distributions in time series models. The main arguments put forward by Phillips refer to the inadequacy of flat priors to reflect complete ignorance about the parameters in a time series setting. In a linear fixed regressors regression context, conditioning on the data is completely neutral, since the parameters to be estimated (the first order ones) do not affect the second moment of the dependent variable. The "neutrality" of flat priors, Phillips argues, is in this case reflected by the fact that the *HPD* confidence sets obtained are analogous to the ones coming from the application of the usual sampling theory techniques.

When we deal with linear time series models, the AR ones for example, as Phillips argues, things appear to change radically, in that the parameters directly affect the correlation structure of the data: conditioning on the sample moment matrix of the data with a flat prior on the parameters is no longer innocuous. Therefore, firstly, a flat prior should be considered as highly informative. Secondly, adopting it induces serious bias in the posterior pdf towards stationarity and a false impression of precision in inference. For these reasons, Phillips suggests that a true ignorance prior should be used, in order to obtain correct neutral posterior distributions.

The necessity of having sharp rules for the specification of the prior distribution intended to reflect complete ignorance has been felt to be particularly urgent, given the need to be able to provide the classical sampling theory results with a Bayesian interpretation. As we have seen in Section [3.3], Jeffreys (1961) indicated that the prior distribution should be chosen to be proportional to the square root of the determinant of the information matrix. This rule has been justified in a variety of ways. Firstly, it minimizes the amount of extra-data information (Lindley, 1961). Secondly, it has nice invariant properties, such as with respect to one-to-one transformations of the parameter space, restrictions on it, and substitution of the data with a sufficient set of statistics. For a full account, see Zellner (1971, appendix to chapter 2). Thirdly, as Perks (1947) points out, it should reflect, as a formalisation of the amount of non-sample information available, the anticipated volume of confidence sets. Since their volume is asymptotically proportional to the inverse of the information matrix, under regularity conditions for the likelihood function, a prior of this form reflects the state of these expectations.

In a simple AR(1) model, the likelihood function appears to be more and more concave in the neighbourhood of  $\hat{\rho}$  as the 'true'  $\rho$  gets larger in absolute value. It is not possible to neglect this kind of information, which is available before any data are actually observed: the higher  $|\rho|$ , the greater the amount of information conveyed by the data about it. For this reason, Phillips argues, a flat prior on  $\rho$

should be regarded as informative, in the sense that it overweights low values of its modulus.

In his paper Phillips considers three different models:

$$y_t = \rho y_{t-1} + e_t \quad (109)$$

$$y_t = \alpha + \beta t + \rho y_{t-1} + e_t, \quad (110)$$

$$y_t = \alpha + \beta t + \rho y_{t-1} + \rho^{**}(L) e_t. \quad (111)$$

Note that, when the deterministic trend is present, the parameterisation chosen is the linear one. Therefore the already mentioned problems of interpretation arise, when  $\rho=1$ .

From the specification of the Jeffreys priors in the three cases, Phillips obtains the posterior distributions for  $\rho$  over the unrestricted range  $(-\infty, +\infty)$ . Some simulations are conducted by Phillips, on the basis of data generating mechanisms corresponding to the three models with  $\rho=1$ . The resulting marginal posterior distributions are either unimodal and skewed to the right, or bimodal. In this latter case the resulting highest posterior density confidence sets can be disconnected, precisely as happens in the application of sampling theory techniques.

As seen above, Phillips' analysis only aims at showing how the flat prior approach tends to over-estimate the posterior probability of  $\rho$  lying in the stationarity region. The posterior odds ratio testing framework represents a different setting, where the unit root hypothesis is compared and tested against an alternative defined over the stationarity interval, and explosive behaviour is ruled out. Clearly, the testing outcomes are bound to depend on the prior distributions which are assumed to reflect prior information.

The technique devised by Schotman and van Dijk relies on the specification of flat priors, on the basis of a declared stance towards neutrality in the decision problem being faced. We have however to recall that a proper prior pdf is required when

comparing a sharp point hypothesis with a composite alternative. Therefore, in order to use the posterior odds testing framework in the unit root analysis, the use of Jeffreys prior for the parameter embedding the unit root hypothesis is not possible. The most sensible way of proceeding in the absence of strong a priori information is to implement different priors and check whether results tend to be robust or not. Geweke (1994) proceeds in this direction. He considers the parameterisation used by Schotman and van Dijk (1991b), and specifies a prior distribution of the kind:

$$p(\rho|H_1) = (s+1) \rho^s I_{[0,1)}(s). \quad (112)$$

The hyperparameter  $s$  is meant to be chosen according to the sampling frequency of the observed data. In Geweke's paper posterior odds ratios are evaluated for different values of  $s$ .

In short, special care should be taken in the specification of prior pdf's in the absence of prior information. The construction of posterior odds ratios to compare a point hypothesis to a composite one limits the range of usable priors, ruling out improper and flat priors over too wide a support. For this reason, it is necessary to specify for the parameters of interest proper prior distributions, parameterised in terms of a manageable set of hyperparameters. Monitoring the effect of different values for the hyperparameters gives a way of clarifying the sensitivity of the posterior analysis.

## **Chapter 4: Bayesian Analysis of Integration at Different Frequencies in Quarterly Data**

### **[4.0] An Overview of the Chapter.**

This chapter provides a unifying framework for conducting Bayesian inference on the presence of seasonal and zero frequency unit roots in quarterly data. The main technique used is the analysis of posterior odds ratios. A new parameterisation is provided for the model, and the prior distributions implemented are discussed and justified. The analysis relies heavily on the application of a Gibbs sampling algorithm. The methods are applied to a set of UK quarterly series. Compared to previous studies, less evidence is found to support seasonal integration hypotheses. The motivations of the study described in this chapter are related to the difficulties of the classical inference approach to unit root testing discussed in details in Chapter 2, and can be summarised as follows. First, unit root test statistics generally have non-standard asymptotic distributions under the null and the associated critical values have to be obtained numerically. Thus there exist a distributional asymmetry between the two hypotheses considered. Second, the Neyman-Pearson inferential apparatus assigns asymmetrical roles to the two hypotheses. Third, all unit root tests share the alarming feature of having unsatisfactory power properties.

In the Bayesian framework described in Chapter 3, it is possible to devise inferential strategies with properties that diverge substantially from those of the classical techniques. In the posterior odds ratio inference setting, the hypotheses being compared are treated in a symmetric fashion, their relative plausibility being gauged on the basis of the corresponding posterior probability. Testing is fully



"consistent", in that the probability of picking the wrong model goes to zero as the sample size increases.

The more general advantages and disadvantages of the Bayesian techniques also apply to the problem of detecting seasonal unit roots. The advantages, together with the ones already described, are essentially related to the fact that the Bayesian approach is simple, and it constitutes the only logical formalization of the process of learning. The disadvantages are mainly of a computational kind. Bayesian methods are "labour intensive techniques". In the present chapter, a Bayesian procedure is applied to testing the seasonal features of quarterly data. The effects of blind reliance on published seasonally adjusted series are well known (see Wallis, 1974). In particular the adjustment of data prior to modelling might result in biased inference. Therefore, it is important to analyse the seasonal features of quarterly time series on a univariate basis.

Given that the classical seasonal unit roots tests (like Hylleberg, Engle, Granger and Yoo (1990), henceforth *HEGY*) appear to share the poor properties of the zero frequency unit root test, I take the less travelled road, and explore the application of Bayesian inference techniques.

The chapter begins by devising a parameterisation which seems to be particularly well suited for the inferential setting being proposed. Section [4.1] is devoted to describing the general characteristics of the *AR* model used, together with the particular parameterisation chosen. The specification allows easy discrimination between deterministic and stochastic seasonality (in the form of the occurrence of seasonal unit roots), and between trend and difference stationarity. In section [4.2], the structure of the prior pdfs is presented, while section [4.3] describes how the joint posterior is dealt with via application of a Gibbs sampling scheme. Section [4.4] is ancillary to this, since it presents the descriptions of the conditional posterior distribution for subsets of the parameter vector. Section [4.5] provides a

description of the posterior odds ratio intended to ease the computing burden. Section [4.6] contains the results of the application conducted on a set of UK quarterly series, and section [4.7] concludes. Appendices [4.A], [4.B], and [4.C] deal with the strict technicalities of the analysis. Appendix [4.A] contains the detailed description of the conditional posterior distributions of different groups of parameters. Appendix [4.B] illustrates the key aspects of the rejection sampling algorithms being implemented. Appendix [4.C] contains the proofs of the smooth transition results.

#### [4.1] General Features of the Model

I consider an autoregressive model for quarterly data. Seasonality can be either deterministic or stochastic. Deterministic seasonality can be accounted for by introduction of dummy variables. Stochastic seasonality requires the application of an adequate filter to induce stationarity and raises the issue of the occurrence of seasonal cointegration (see *HEGY*, 1990 and Engle et al., 1993). We have stochastic seasonality when the *AR* polynomial contains some unit modulus roots at seasonal frequencies. In the quarterly case, the seasonal roots are -1, for the biannual cycle, and  $\pm i$ , for the annual cycle.

The model considered for the observable variable  $z_t$  is the following:

$$\phi(L) y_t = e_t, e_t \sim N.i.d(0, \sigma^2), \quad (1)$$

$$y_t = z_t - S_t - \eta_t, \quad (2)$$

$$S_t = \alpha_0 + \alpha_1 \cos[(\pi/2)t] + \beta_1 \sin[(\pi/2)t] + \alpha_2 \cos[\pi t] \quad (3)$$

$$\phi(L) = 1 - \phi_1 L - \phi_2 L^2 - \dots - \phi_k L^k. \quad (4)$$

The hypotheses of interest concern the roots of the equation  $\phi(L) = 0$ , and are as follows:

- 1)  $\phi(-1) = 0$  (integration at frequency  $\lambda = \pi$ : semi-annual cycle)
- 2)  $\phi(i) = \phi(-i) = 0$  (integration at frequency  $\lambda = \pi/2$ : annual cycle)
- 3)  $\phi(1) = 0$  (integration at frequency  $\lambda = 0$ : the series is difference stationary)
- 4) Any combination of the above hypotheses.

Each one of the "null" hypotheses considered is compared to a parallel "alternative", in which the envisaged non-stationary feature is modelled with an appropriate deterministic component. The last part of this section is devoted to setting out the chosen parameterisation. In the next section the structure of the prior used is outlined and justified.

The model is cast in terms of the parameterisation used by *HEGY*, which makes use of the Laplace expansion of  $\phi(L)$  around the roots  $\pm 1$ , and  $\pm i$ :

$$\phi^*(L)y_{4t} = \psi_1 y_{1t-1} + \psi_2 y_{2t-1} + \psi_3 y_{3t-2} + \psi_4 y_{3t-1} + e_t, \quad (5)$$

in which  $\phi^*(L)$  is a polynomial in  $L$  with degree  $k^* = k - 4$ ,  $\psi_1$ ,  $\psi_2$ ,  $\psi_3$  and  $\psi_4$  are linear functions of the parameters in  $\phi(L)$ , and the variables  $y_{1t}$ ,  $y_{2t}$ ,  $y_{3t}$  and  $y_{4t}$  are defined in the following terms:

$$\begin{aligned} y_{1t} &= (1 + L + L^2 + L^3) y_t, \\ y_{2t} &= -(1 - L + L^2 - L^3) y_t, \\ y_{3t} &= -(1 - L^2) y_t, \\ y_{4t} &= (1 - L^4) y_t = (1 - L) S(L) y_t, \quad S(L) = 1 + L + L^2 + L^3. \end{aligned} \quad (6)$$

In *HEGY's* setting the hypotheses of integration at different frequencies are represented as the following restrictions on the representation :

frequency	$\lambda = 0:$	$\psi_1 = 0 ;$
frequency	$\lambda = \pi/2:$	$\psi_3 = \psi_4 = 0 ;$
frequency	$\lambda = \pi:$	$\psi_2 = 0$

We consider now how to use Bayesian inference techniques for the analysis of such hypotheses. In order to ease the implementation, a variant of the parameterisation (5) is used, so as to represent the  $\pi/2$  integration hypothesis as a restriction on a single parameter. In fact defining :

$$\psi_3 = -2r \cos \theta, \quad \psi_4 = 2r \sin \theta, \quad (7)$$

I can write the model as:

$$\phi^*(L)y_{4t} = \psi_1 y_{1t-1} + \psi_2 y_{2t-1} + 2r (\sin \theta y_{3t-1} - \cos \theta y_{3t-2}) = e_t, \quad (8)$$

or equivalently (since  $y_t = z_t' S_t - \gamma t$ ):

$$\begin{aligned} \phi^*(L)z_{4t} - \psi_1 z_{1t-1} - \psi_2 z_{2t-1} - 2r (\sin \theta z_{3t-1} - \cos \theta z_{3t-2}) = \\ = [4\phi^*(1) + 10\psi_1 + 2\psi_2 + 4r (\sin \theta - \cos \theta) - 4\psi_1 t] \gamma - 4\psi_1 \alpha_0 \\ + 4r \cos [(\pi/2)t - \theta] \alpha_1 + 4r \sin [(\pi/2)t - \theta] \beta_1 - 4\psi_2 \cos (\pi t) \alpha_2 + e_t, \end{aligned} \quad (9)$$

where:

$$\begin{aligned} z_{1t} &= (1+L+L^2+L^3) z_t, \\ z_{2t} &= -(1-L+L^2-L^3) z_t, \end{aligned} \quad (10)$$

$$z_{3t} = -(1-L^2) z_t,$$

$$z_{4t} = (1-L^4) z_t = (1-L) S(L) z_t.$$

We have integration at frequency  $\pi/2$  when  $r = 0$ ; in such an occurrence, the parameters  $\alpha_1, \beta_1$  and  $\theta$  disappear. When there is integration at frequency zero, the parameter  $\alpha_0$  disappears, and so does the trend term  $-4\psi_1 \gamma t$ . The model is difference stationary. Under the hypothesis of integration at frequency  $\pi$ , the parameter  $\alpha_2$  disappears.

It is evident that the model is not linear in the parameters involved. Nevertheless, I believe that it may provide a sensible framework to conduct inference, because it is based on a "structural" parameterisation (see Barghava, 1986, Schmidt and Phillips, 1992). No parameter is redundant under any of the hypotheses considered. A linear parameterisation of the form:

$$\phi(L) y_t = \text{deterministics} + e_t, \quad (11)$$

would be analytically much more manageable but would not allow the different integration hypotheses to be as neatly defined as in the model with the "structural" parameterisation.

The specification of prior distributions is bound to yield analytically intractable posterior distributions, since model (12) does not allow any natural-conjugate analysis. For this reason, in order to deal with the posterior distribution it is necessary to resort to simulation methods, such as Markov Chain Monte Carlo integration.

#### [4.2] The Specification of the Priors

In order to describe the application of the simulation techniques used in this chapter, the parameters of the model can be divided in 7 different groups:  $\eta = [\eta_1, \eta_2, \eta_3, \eta_4, \eta_5, \eta_6, \eta_7]'$ , where:

$$\begin{aligned}\eta_1 &= \beta^* = [\beta' \gamma']', \beta = [\alpha_0 \alpha_2 \alpha_1 \beta_1]', \\ \eta_2 &= [\phi^*]', \phi^* = [\phi_1^* \dots \phi_k^*]', \\ \eta_3 &= \sigma, \eta_4 = \psi_1, \eta_5 = \psi_2, \eta_6 = r, \eta_7 = \theta.\end{aligned}\tag{12}$$

As it will become clear in the next sections, this division is done in order to associate these subsets of parameters with tractable conditional posterior distributions.

We also adopt the notation  $\bar{\eta}_i, i = 1, \dots, 7$ , to indicate that subset of parameters in  $\eta$  such that  $\bar{\eta}_i \cup \eta_i = \eta$ .

The following prior distribution structure is put forward:

$$\begin{aligned}p(\beta^*, \gamma | \bar{\eta}_1) &\sim N(b^*, \sigma^2 V^*), \\ V^* &= \text{diag}(-\psi_1^{-1}, -\psi_2^{-1}, r^{-1}, r^{-1}, (\sigma_\gamma/\sigma)^2), b^* = [b', \mu_\gamma]', b = [a_0, a_2, a_1, b_1]', \\ p(\phi^*) &\propto 1, \phi^* \in \mathbb{R}^k \\ p(\sigma) &\propto \sigma^{-1}, \sigma \in \mathbb{R}_+, \\ p(\psi_i) &= \lambda_i \exp(\lambda_i \psi_i), \psi_i \in \mathbb{R}_+, i = 1, 2; \\ p(r) &= \lambda_r \exp(-\lambda_r r), r \in \mathbb{R}_+; \\ p(\theta) &\sim N[\mu_\theta, \sigma_\theta], \theta \in [-\pi/2, +\pi/2].\end{aligned}\tag{13}$$

The choice of the priors is justified in the following way:

(1) The prior on  $\beta$ , the parameters of the deterministic seasonal structure, is 4-variate normal, around a location vector  $b$  which is determined on the basis of the initial observations of the process. The prior variances of the single elements of  $\beta$  are designed to go to infinity as the model approaches the corresponding frequency integration setting. We have that:

$$\lim_{\psi_1 \rightarrow 0} V_{11} = \infty, \quad \lim_{\psi_2 \rightarrow 0} V_{22} = \infty, \quad \lim_{r \rightarrow 0} V_{ii} = \infty, \quad i = 3, 4, \quad (14)$$

$$\lim_{\psi_1 \rightarrow 0} [\psi_1^2 V_{11}] = 0, \quad \lim_{\psi_2 \rightarrow 0} [\psi_2^2 V_{22}] = 0, \quad \lim_{r \rightarrow 0} [r^2 V_{ii}] = 0, \quad i = 3, 4, \quad (15)$$

i.e. the prior precisions go to zero, but slower than  $\psi_1^2$ ,  $\psi_2^2$  and  $r^2$ . This property is particularly important because it ensures that the deterministic component of the "reduced form" model has a logically sound prior distribution, and that the posterior distribution of the parameters under the stationary alternative passes smoothly to the posterior distribution under the different integration hypotheses being considered. The analytical proofs of the smooth transition properties are in Appendix [4.C].

The linear trend parameter  $\gamma$  is given a normal prior, with position and variance specified by means of the two corresponding hyperparameters  $(\mu_\gamma, \sigma_\gamma)$ . Of course the choice of such hyperparameters is entirely subjective. The specification of a flat prior for  $\gamma$  would induce only marginal modifications to our analysis, and can be seen as a particular case, when the prior precision goes to zero.

2) The parameters in the transient AR dynamics, i.e. on the lags of  $y_{4t}$ , are given a flat prior, just to ease the computations. Of course the specification of more articulated priors is possible. For instance, as in Geweke (1994), one could put a prior on each of them along the lines of Doan, Litterman and Sims (1984): normal prior with zero mean and prior variance that shrinks to zero as the lag order

increases. This would not modify our analysis very much. It nevertheless appears that I do not have their problems of overparameterisation here, and I can focus on models with not too many lags.

(3) The prior on the variance parameter  $\sigma$  is customarily a Jeffreys prior.

(4) and (5) As for the parameters  $\psi_1$  and  $\psi_2$ , which are associated with the hypotheses of zero frequency and  $\pi$  frequency integration, I choose to specify negative-exponential priors with hyperparameters  $\lambda_1$  and  $\lambda_2$ . It is believed that such a functional form is quite appropriate because:

i) it does not force any restriction on the support of the parameters, other than the legitimate one of stationarity.

ii) it is a proper non-flat prior which therefore can be used in a posterior odds ratio testing framework involving sharp point nulls. The problems outlined in Schotman and van Dijk (1991a, 1991b, 1992) and regarded as occurrences of the 'Lindley paradox' (see Lindley, 1957) in the specification of flat priors under a composite hypothesis are therefore solved. The researcher has to provide a choice for the hyperparameters involved. Since these two parameters are the inverse of the respective prior means, in the absence of any extra-sample observation, one might choose them to be equal to the inverse of the unrestricted *OLS* estimates of the *HEGY* parameterisation.

(6) For  $r$  exactly the same kind of prior is chosen, except that the parameter  $r$  is defined to be positive.

(7) The prior on  $\theta$ , the phase angle in the deterministic  $\pi/2$  seasonal, poses some problems. Although the parameter  $\theta$  ceases to be identified under the  $\pi/2$  frequency unit root hypothesis, it is not possible to assign it a prior dispersion determined on the basis of  $r$ . This would put obstacles in the way of the smooth transition results, which form the basis of the posterior odds ratio evaluation. Therefore I use a truncated normal distribution, with support  $[-\pi/2, \pi/2]$ . Other



choices, such as Cauchy or Student- $t$ , are equally legitimate. The choice of the hyperparameters is based on the *OLS* estimates of  $\psi_3$  and  $\psi_4$  in the standard *HEGY* parameterisation. Of course it is necessary to provide a check for sensitivity with respect to the specification of all the prior distributions. This comes with the results of the application contained in section [4.6].

#### [4.3] The Joint Posterior Distribution

The likelihood of the model can be written as follows:

$$p(\text{data} \mid \eta \mathbf{D}_0) \propto \sigma^{-T} \exp\left\{-\frac{1}{2\sigma^2} \mathbf{e}'\mathbf{e}\right\}, \quad \mathbf{e} = \{e_t\}_{t=1}^T. \quad (16)$$

In the text the notation " $\mathbf{D}_t$ " means conditional on the data evidence up to period  $t$ , and therefore when I condition upon  $\mathbf{D}_0$  or  $\mathbf{D}_T$ , I respectively indicate "conditional on initial conditions" or "conditional on the whole sample information" (a posteriori). Combining the information provided by the prior distribution with the likelihood function, I obtain the joint posterior:

$$\begin{aligned} p(\eta \mid \mathbf{D}_t) &\propto \sigma^{-(T+5)} (-\psi_1)^{1/2} (-\psi_2)^{1/2} r \exp\left\{-\frac{1}{2\sigma^2} [\mathbf{E}'\mathbf{E} + \sigma_\theta^2(\theta - \mu_\theta)^2]\right. \\ &\quad \left.+ \lambda_1 \psi_1 + \lambda_2 \psi_2 - \lambda_T r\right\}, \\ \mathbf{E} &= \begin{bmatrix} \mathbf{e} \\ \mathbf{V}^{*-1/2}(\boldsymbol{\beta}^* - \mathbf{b}^*) \end{bmatrix}. \end{aligned} \quad (17)$$

When it comes to conducting inference on a subset of parameters of interest, on the basis of the posterior pdf, it is clear that I have to be able to marginalise it with respect to the parameters I am not interested in, i.e. the nuisance parameters.

Expression (17) does not allow the possibility of easily obtaining marginal distributions or posterior moments on the basis of available analytical results. A possibility would be to resort to approximations, such as the ones described in Chapter [3]. But it is difficult to obtain manageable results, and, above all, no information is obtained about the reliability of these approximations. On the other hand, numerical integration is not feasible, given the high dimensionality of the parameter space (I have in total  $k+6$  parameters). One has then to rely on fast, efficient and precise numerical simulation techniques.

Suppose that our interest focused on the posterior mean of some function of the parameters, say  $f(\boldsymbol{\eta})$ :

$$E(f(\boldsymbol{\eta})|\mathbf{D}_T) = \int f(\boldsymbol{\eta}) p(\boldsymbol{\eta}|\mathbf{D}_T) d\boldsymbol{\eta} . \quad (18)$$

In section [4.5], I show how the posterior odds ratio (henceforth *POR*) can be thought of as the posterior expectation of a particular function of certain parameters. The posterior moments can be computed numerically to an arbitrary degree of accuracy on the basis of the Monte Carlo integration principle. If it were possible to obtain  $N$  draws from the joint posterior pdf, say  $\boldsymbol{\eta}^{(i)}$ ,  $i=1, 2, \dots, N$ , then the posterior expectation of  $f(\boldsymbol{\eta})$  could be easily estimated as the sample mean:

$$\bar{f}_N = N^{-1} \sum_{i=1}^N f(\boldsymbol{\eta}^{(i)}) \quad (19)$$

Given an *i.i.d.* assumption on the draws, the law of large numbers ensures convergence of the above expression to the posterior expectation of  $f(\boldsymbol{\eta})$ . Of course as  $N$  increases, so does the accuracy of the estimate.

If it is not possible to provide *i.i.d.* draws from the joint posterior distribution, as in our case where it is of no known analytical form, then some other methods have to

be adopted. Following Geweke (1989), one could choose an "importance function", to sample from. As discussed in Chapter 3, that choice is not easy, and it might yield very poor estimates, especially when the shape of the posterior distribution is not known in all its details. We do not know the form of the joint posterior in our context, and therefore I adopt a Markov Chain sampling scheme, and more precisely a Gibbs Sampling scheme (*GSS*). The idea behind the *GSS* is quite simple and intuitive, as described in detail in Chapter 3. All that is needed is that the conditional posterior distributions of a class of mutually exclusive exhaustive subsets of the parameters  $p(\eta_i | \bar{\eta}_i, \mathbf{D}_T)$  be "available" in the sense that each of them can be easily simulated. If these conditions are met the *GSS* works as follows. We start from an arbitrary initialisation of the parameter vector:

$$\boldsymbol{\eta}^{(0)} = [\eta_1^{(0)}, \eta_2^{(0)}, \dots, \eta_k^{(0)}]' \quad (20)$$

At each pass of the algorithm, a random draw from each of the  $p(\eta_i | \bar{\eta}_i, \mathbf{D}_T)$  distributions is obtained, and the results from the draw are used to condition the posterior distributions in the next pass. Hence a Markovian updating scheme is obtained, in which the draws are not independent, nor identically distributed. In Chapter 3, I have listed some sufficient conditions to ensure that: (1) the continuous state Markov chain induced by the *GSS* converges in distribution to the true joint posterior distribution at a rate which is geometric in the number of passes used in the algorithm. (2) The numerical estimate of the posterior mean of any function of the parameters (if it exists) converges a.s. to its true value. What has to be shown is how to characterise the conditional posterior pdfs in our context, to check whether they comply with the conditions required for convergence to hold, and how it is possible to obtain random drawings from them. This is the object of the next section.

#### [4.4] The conditional posterior distributions

Even if the joint posterior distribution has no known analytical form, it turns out to be possible to characterize the conditional posterior distributions of some subsets of parameters. Some of these pdf's are of known analytical form, whereas others are not; in Appendix [4.B], I describe how it is possible to apply the method of rejection sampling to the conditional posterior pdf's whose analytical form is non-standard.

Starting from expression (17) I readily obtain some results for the 7 different groups of parameters in the model. These results are presented as lemmata, whose proofs are contained in Appendix [4.A].

##### Lemma 4.4.1

$p(\eta_1 | \overline{\eta_1} \mathbf{D}_T)$ , where  $\eta_1 = [\beta' \gamma]' = \beta^*$ , is 5-variate normal, from which independent random draws are readily obtained.

##### Lemma 4.4.2

$p(\eta_2 | \overline{\eta_2} \mathbf{D}_T)$  is  $k^*$ -variate normal, and again independent random draws are readily obtained. Remember that  $\eta_2 = [\phi^*]$  is a  $(k^* \times 1)$  vector containing the parameters on the lags of  $y_{4t} = z_{4t} - 4\gamma$ .

##### Lemma 4.4.3

$p(\eta_3 | \overline{\eta_3} \mathbf{D}_T)$  where  $\eta_3 = \sigma$ , allows for indirect drawing through a  $\chi^2$  distribution.

##### Lemma 4.4.4

$p(\eta_4 | \overline{\eta_4} \mathbf{D}_T)$ , with  $\eta_4 = \psi_1$ ,  $p(\eta_5 | \overline{\eta_5} \mathbf{D}_T)$ , with  $\eta_5 = \psi_2$ ,  $p(\eta_6 | \overline{\eta_6} \mathbf{D}_T)$ , with  $\eta_6 = r$ , and  $p(\eta_7 | \overline{\eta_7} \mathbf{D}_T)$ , with  $\eta_7 = \theta$ , have no standard form. Their simulations require

rejection sampling. Appendix [4.B] contains a complete account of the choices made in terms of functional form for the reference distributions and their parameters.

By inspection of the conditional posterior pdf's, it is immediate to verify that the sufficient conditions for convergence of the *GSS* are fulfilled. Remember that these conditions are:

- (1)  $p_j(\eta_j | \eta_i, i \neq j) > 0 \forall \eta_i \in H_i, \eta_j \in H_j$ ,
- (2)  $P_j(A | \eta_i, i \neq j) > 0 \forall P_j$ -measurable set  $A \subseteq H_j$ .
- (3)  $P_j(A | \eta_i, i \neq j)$  is a continuous function of  $\eta_i, i \neq j$ ,

where  $H = [H_1 \times H_2 \times \dots \times H_6]$  is the support of the joint posterior pdf.

To summarize, the last two sections show how the resulting joint posterior distribution is not of any analytically known kind. Nevertheless, given that we are able to draw independently from an exhaustive set of conditional posterior pdf's, and putting these draws into a Markov chain sequence, I can apply a *GSS* to obtain synthetic draws from the joint posterior pdf. These draws form the basis for the evaluation of the posterior moments of any function of the parameters. We next show that the posterior odds ratios of hypotheses can be seen as posterior means of certain functions of the parameters.

#### [4.5] A Convenient Description of the Posterior Odds Ratio

In the particular context described in this section, the *POR* can be written in a way that is very convenient for computations. Suppose we are interested in comparing two competing hypotheses concerning  $\eta$ , the parameter vector of a certain model:

$$H_A: p_A(\eta), \eta \in \Theta_A,$$

$$H_B: p_B(\eta), \eta \in \Theta_B,$$

where  $\Theta_A$ , and  $\Theta_B$  are the supports of  $\eta$  under  $H_A$  and  $H_B$  respectively. Associated with the two hypotheses I have, as usual, two families of priors,  $p_A(\eta)$  and  $p_B(\eta)$ .

The *POR* is usually defined as:

$$POR = \frac{p(H_A|data)}{p(H_B|data)} = \frac{\int_{\Theta_A} p_A(\eta) p(data|\eta, H_A) d\eta}{\int_{\Theta_B} p_B(\eta) p(data|\eta, H_B) d\eta} \quad (21)$$

In this case two alternative hypotheses for the same model with likelihood function  $L(\eta)$  are contemplated. In other words,  $L(\eta) = p(data|\eta, H_A) = p(data|\eta, H_B)$ . This does not mean that in this case the data are not informative on the hypotheses being compared<sup>1</sup>, but that the model has the same likelihood function under both hypotheses.

Therefore, it is possible to write (see Geweke, 1994, p.619):

$$POR = \frac{\int_{\Theta_A} \frac{p_A(\eta)}{p_B(\eta)} p_B(\eta) L(\eta) d\eta}{p(data|H_B)} = \int_{\Theta_A} \frac{p_A(\eta)}{p_B(\eta)} p(\eta|D_T, H_B) d\eta \quad (22)$$

Expression above states that the *POR* can be obtained by averaging over  $\Theta_A$  the function:

$$f(\eta) = p_A(\eta)/p_B(\eta). \quad (23)$$

<sup>1</sup>This would occur if  $p(data|H_A) = p(data|H_B)$ .

using the posterior distribution  $p(\eta|\mathbf{D}_T, H_B)$  as weighting function.

The context is only slightly more complicated when comparing a point hypothesis against a composite competing one. As a simple example of this, let us consider the partition  $\eta = [\eta_1', \eta_2']$ , and the hypotheses

$$H_A: \eta_1 = \mathbf{h}_1, \eta_2 \in \Theta_2,$$

$$H_B: \eta_1 \in \Theta_1, \eta_2 \in \Theta_2,$$

with prior pdf:  $p(\eta_1 = \mathbf{h}_1|H_A) = 1, p(\eta_1|H_B), p(\eta_2|H_A) = p(\eta_2|H_B)$ .

In this case expression (22) becomes the so-called "Savage density ratio"<sup>2</sup> (Kass and Raftery, 1995, p. 780):

$$POR = \frac{1}{p(\eta_1 = \mathbf{h}_1|H_B)} \int_{\Theta_2} p(\eta_1 = \mathbf{h}_1, \eta_2|\mathbf{D}_T, H_B) d\eta_2 = \frac{p(\eta_1 = \mathbf{h}_1|\mathbf{D}_T, H_B)}{p(\eta_1 = \mathbf{h}_1|H_B)}, \quad (24)$$

given that Dickey (1971, pp. 208-9) attributes this result to L. J. Savage.

When the joint posterior distribution under  $H_B$  is analytically intractable, a convenient way to calculate the numerator in the expression above is to evaluate analytically the conditional posterior pdf  $p(\eta_1 = \mathbf{h}_1|\eta_2, \mathbf{D}_T, H_B)$ , and average it over draws taken from the posterior distribution  $p(\eta_2|\mathbf{D}_T, H_B)$ . The *POR* is hence consistently estimated via Monte Carlo simulation as:

$$\overline{POR} = \frac{1}{p(\eta_1 = \mathbf{h}_1|H_B)} \times \frac{1}{N} \sum_{i=1}^N p(\eta_1 = \mathbf{h}_1|\eta_2^{(i)}, \mathbf{D}_T, H_B). \quad (25)$$

In the applications presented in this chapter, I am interested in gauging the posterior evidence in support of the presence of different unit moduli roots. The

<sup>2</sup> Dickey (1971) attributes this result to L.J. Savage.

sharp point hypothesis can consider just one unit root at a time or more, and it is compared directly to the stationary alternative specification.

In Section [4.1] I showed how each point hypothesis induces a corresponding subset of the parameters to become unidentified. As for the computation of *POR*'s, this circumstance induces some complications which can be tackled as follows. For explanatory purposes, I restrict attention to the  $\pi/2$  frequency integration hypothesis, and I consider a comparison between  $H_A: r = 0$  and  $H_B: r > 0$ . Comparisons involving different frequency integration hypotheses (even between joint hypotheses) can be conceptually dealt with on the basis of exactly the same framework.

We consider  $H_A$  as the limiting expression, for  $\varepsilon$  approaching zero, of the following hypothesis:

$$\begin{aligned} H_A(\eta): r &\in (0, \varepsilon), \\ p_A(\eta) &= \varepsilon^{-1} I_{(0, \varepsilon)}(r), \quad r \in (0, \varepsilon), \end{aligned} \quad (26)$$

with  $I_{(0, \varepsilon)}(r)$  the indicator function, taking unit values within the  $(0, \varepsilon)$  interval and equal to zero elsewhere.

For homogeneity, I restore the notation used in Section [4.2] and indicate  $r$  as  $\eta_6$ ; all the other parameters in the model are collected in the set  $\bar{\eta}_6$ .

For  $\bar{\eta}_6$  I adopt the prior pdf specification discussed in section [4.2], i.e.  $p_A(\bar{\eta}_6 | \eta_6) = p_B(\bar{\eta}_6 | \eta_6)$ . As for  $\eta_6$ , I specify the same prior as in section [4.2] under  $H_B$ ; under  $H_A$ , a flat prior is adopted as described in expression (26).

I can therefore write the *POR* as follows:

$$POR = \int \left[ \frac{\int_0^\varepsilon \frac{p_A(\eta_6)}{p_B(\eta_6)} p(\eta_6 | \bar{\eta}_6, \mathbf{D}_T, H_B) d\eta_6 \right] p(\bar{\eta}_6 | \mathbf{D}_T, H_B) d\bar{\eta}_6. \quad (27)$$



The posterior expectation of the function  $f(\eta_6) = p_A(\eta_6)/p_B(\eta_6)$ , conditional on the other parameters in  $\bar{\eta}_6$ , has to be averaged over  $(0, \varepsilon)$  by using the posterior pdf of  $\bar{\eta}_6$  as a weighting function. From the discussion in section [4.2], I already know the form of the conditional posterior distribution of  $\eta_6$  (see also Appendix [4.B]). In what follows, I show what happens when  $\varepsilon \rightarrow 0$ . I make use of the smooth transition results, as detailed in Appendix [4.C].

It is easy to see that the function of interest, i.e.:

$$f(\eta_6) = \varepsilon^{-1} I_{(0, \varepsilon)}(r) \lambda_r^{-1} \exp(\lambda_r r), \quad (28)$$

only depends on  $\eta_6$ . Theoretically one could marginalise out all the parameters but  $\eta_6$ , and evaluate the *POR* as posterior expectation of  $f(\eta_6)$  on the basis of the uni-dimensional posterior pdf of  $\eta_6$ . That is analytically impossible. In order to make efficient use of the numerical evaluation techniques being used, I can marginalise with respect to the parameters of the deterministic representation that disappear when  $r = 0$ , namely  $\alpha_1$  and  $\beta_1$ .

I define  $\bar{\eta}_6^*$  such that :

$$\bar{\eta}_6^* = (\alpha_1, \beta_1, r), \quad (29)$$

i.e.  $\bar{\eta}_6^*$  is the vector of all the parameters, bar  $\alpha_1$ ,  $\beta_1$ , and  $r$ . The *POR* is then:

$$POR = \int \left[ \int_0^\varepsilon \frac{p_A(\eta_6)}{p_B(\eta_6)} p(\eta_6 | \bar{\eta}_6^*, \mathbf{D}_T, H_B) d\eta_6 \right] p(\bar{\eta}_6^* | \mathbf{D}_T, H_B) d\bar{\eta}_6^* \quad (30)$$

The results in Appendix [4.C] then allow us to write:

$$p(\eta_6 | \bar{\eta}_6^*, \mathbf{D}_T, H_B) = [1 + 8 T r]^{-1} \exp \left\{ -\frac{1}{2\sigma^2} \mathbf{w}^{*'} \mathbf{M}(\mathbf{X}^*) \mathbf{w}^* - \lambda, r \right\} / k(\bar{\eta}_6^*), \quad (31)$$

$$k(\bar{\eta}_6^*) = \int_0^\infty [1 + 8 T r]^{-1} \exp \left\{ -\frac{1}{2\sigma^2} \mathbf{w}^{*'} \mathbf{M}(\mathbf{X}^*) \mathbf{w}^* - \lambda, r \right\} d\eta_6,$$

where the  $[(T+2) \times 1]$  vector  $\mathbf{w}^*$  and the  $[(T+2) \times 2]$  matrix  $\mathbf{X}^*$  are defined as:

$$\begin{aligned} \mathbf{w}^* &= [r^{1/2} a_1, r^{1/2} b_1, \mathbf{w}']', \quad \mathbf{X}^{*'} = [r^{1/2} \mathbf{I}_2, \mathbf{X}']', \quad \mathbf{w} = \{w_t\}_{t=1}^T, \quad \mathbf{X} = \{\mathbf{x}_t\}_{t=1}^T, \\ w_t &= \phi^*(L) y_{4t} - \psi_1 y_{1t-1} - \psi_2 y_{2t-1} - 2r [( \sin \theta )(z_{3t-1} + 2\gamma) - ( \cos \theta )(z_{3t-2} + 2\gamma)], \quad (32) \\ \mathbf{x}_t' &= 4r \{ \cos [(\pi/2) t - \theta], \cos [(\pi/2) t - \theta], \mathbf{M}(\mathbf{X}^*) = \mathbf{I}_{T+2} - \mathbf{X}^* (\mathbf{X}^{*'} \mathbf{X}^*)^{-1} \mathbf{X}^{*'} \}. \end{aligned}$$

Therefore the *POR* is:

$$POR = \int \left[ \int_0^\varepsilon \frac{\exp \left\{ -\frac{1}{2\sigma^2} \mathbf{w}^{*'} \mathbf{M}(\mathbf{X}^*) \mathbf{w}^* \right\}}{\varepsilon \lambda, [1 + 8 T r]} dr \right] / k(\bar{\eta}_6^*) \left\{ p(\bar{\eta}_6^* | \mathbf{D}_T, H_B) d\bar{\eta}_6^* \right\}. \quad (33)$$

Given the smooth transition results, as  $\varepsilon$  shrinks to zero and the prior distribution of  $r$  under  $H_A$  attains unit probability mass at  $\eta_6 = r = 0$ , the *POR* becomes:

$$POR = \int \left\{ \left[ \lambda,^{-1} \exp \left\{ -\frac{1}{2\sigma^2} \mathbf{e}_A' \mathbf{e}_A \right\} \right] / k(\bar{\eta}_6^*) \right\} p(\bar{\eta}_6^* | \mathbf{D}_T, H_B) d\bar{\eta}_6^*, \quad (34)$$

where  $\mathbf{e}_A$  is the vector of the error terms in the model under the  $\pi/2$  frequency integration hypothesis:

$$\mathbf{e}_{At} = \phi^*(L) y_{4t} - \psi_1 y_{1t-1} - \psi_2 y_{2t-1}.$$

This means that the function of interest :

$$f(\bar{\eta}_6^*) = \frac{p(\eta_6 = 0 | \mathbf{D}_T, \bar{\eta}_6^*, H_B)}{p(\eta_6 = 0 | H_B)} = \left[ \lambda_r^{-1} \exp \left\{ -\frac{1}{2\sigma^2} \mathbf{e}_A' \mathbf{e}_A \right\} \right] / k(\bar{\eta}_6^*), \quad (35)$$

which depends on all parameters of the model but  $\alpha_1$ ,  $\beta_1$ ,  $r$  and  $\theta$ , is evaluated for any draw from the posterior distribution of  $\bar{\eta}_6^*$ , and then averaged numerically, on the basis of the draws obtained from the posterior distribution under  $H_B$ . The application of the *GSS* renders this approach feasible.

Note that expression (34) is clearly the same as expression (24), the result quoted by Kass and Raftery, since it is a Monte Carlo estimate of  $p(\eta_6=0|\mathbf{D}_T, H_B)/p(\eta_6=0|H_B)$ .

Computing the *POR* this way allows one to avoid using the complicated Monte Carlo procedures described in Newton and Raftery (1994), Gelfand and Dey (1994), and Carlin and Chib (1995). This is due to the nested nature of the hypotheses being compared in the present context, and to the smooth transition results.

The above framework applies to any unit root hypothesis comparison with the associated stationary alternative, and the corresponding *POR* can be obtained exactly in the same conceptual way. For this reason the inferential strategy for the problem under study is as follows: I make use of the Gibbs sampling algorithm to generate draws from the joint posterior distribution under the most general model. At each pass of the sampler I keep track of the relevant functions of the parameters. These functions are averaged to yield the posterior odds ratios.

#### [4.6] An Application

This section presents an application to five of the UK macroeconomic series studied in Osborn (1990), where details of the source of the data can be found. We

consider real GDP, total real consumption (including durables and non-durables), real investment (total gross fixed capital formation), employment, and real narrow money (M0). All the series are in natural logarithms. Data run from 1955, first quarter, to 1988, last quarter, except for M0, for which a shorter sample period is available (1969:3-1988:4).

For these variables, the application of *HEGY*'s testing procedures led to the conclusion that GDP and consumption possess unit roots at all frequencies, whereas investment, employment and real money have only a zero frequency unit root (Osborn, 1990, Table 2). The application of the Bayesian technique only partly confirms these results, as shown below.

Before presenting the results, some explanation of how the univariate models were specified and how prior distribution hyperparameters were chosen is required. The model lag order was chosen on the basis of the application of a series of different criteria: information criteria (Akaike, Hannan and Quinn and Schwartz), variable deletion tests on an over-parameterized general model, and Godfrey portmanteau test to check the validity of the resulting model. It emerged that all the series being analysed required an autoregressive representation of the 5<sup>th</sup> order.

As for the deterministic parameters  $\alpha_0$ ,  $\alpha_2$ ,  $\alpha_1$  and  $\beta_1$ , their respective location hyperparameters  $a_0$ ,  $a_2$ ,  $a_1$  and  $b_1$ , are determined on the basis of the pre-sample observations, treated as initial conditions of the underlying processes. The parameters in the prior distribution of  $\gamma$ ,  $\mu_\gamma$  and  $\sigma_\gamma$  are determined such that  $\mu_\gamma$  match the average in-sample growth rate of the series, and such that the prior distribution assigns 95% of the whole probability mass to the interval  $\mu_\gamma \pm 2\%$ . The hyperparameters of the prior distributions of  $\psi_1$ ,  $\psi_2$ , and  $r$ , i.e. respectively  $\lambda_1$ ,  $\lambda_2$  and  $\lambda_r$ , were determined on the basis of the following procedure. An unrestricted *AR* process was fitted to the data:

$$\phi(L)z_t = \sum_{i=1}^4 \delta_i D_{it} + \zeta t + e_t, \quad (36)$$

and estimated by means of the *OLS* estimator. On the basis of these estimates, indirect estimates for  $\psi_1$ ,  $\psi_2$ , and  $r$  were provided for all the series under study. The reliability of these estimates has been previously gauged on the basis of a Monte Carlo experiment. This experiment points out that these indirect estimates have a well behaved, bell-shaped distribution around the true values. The indirect estimates,  $\tilde{\psi}_1$ ,  $\tilde{\psi}_2$ ,  $\tilde{r}$ , form the basis of the choice of the hyperparameters. In the absence of any a priori information, recall that the parameter of a negative exponential distribution is the reciprocal of its expected value. On the basis of this consideration the hyperparameters were determined as:

$$\lambda_1 = \frac{1}{\tilde{\psi}_1}, \lambda_2 = \frac{1}{\tilde{\psi}_2}, \lambda_r = \frac{1}{\tilde{r}}. \quad (37)$$

While this choice seems plausible and sound, it may have an important bearing on the analysis. Consequently sensitivity analysis of the results with respect to different choices of such hyperparameters is carried out, and the results are discussed at the end of this section.

Finally, as for the hyperparameters of the prior distribution of  $\theta$ ,  $\mu_\theta$  is chosen equal to zero and  $\sigma_\theta$  is determined as that value that gives 95% of the Gaussian probability mass to the  $(-\pi/2, +\pi/2)$  interval.

The hyperparameters used in the application are summarized in Table 4.1, presented here.

Table 4.1: Hyperparameters

	GDP	Cons'n.	Investm.	Employm.	M <sub>0</sub>
$a_0$	10.557	10.190	8.712	10.105	8.208
$a_2$	0.004	0.023	0.046	-0.001	0.016
$a_1$	-0.008	-0.023	0.001	-0.003	-0.020
$b_1$	0.021	0.023	0.019	-0.000	0.003
$\mu_\gamma$	0.006	0.007	0.009	0.001	0.009
$\sigma_\gamma$	1.020	1.020	1.020	1.020	1.020
$\lambda_1$	49.03	39.781	75.803	57.250	69.905
$\lambda_2$	5.915	6.463	5.107	2.749	2.028
$\lambda_\gamma$	5.851	5.436	4.848	3.082	2.859
$\mu_\theta$	0.000	0.000	0.000	0.000	0.000
$\sigma_\theta$	1.603	1.603	1.603	1.603	1.603

On the basis of these hyperparameters, the resulting joint posterior distributions were simulated via application of a Gibbs sampling algorithm<sup>3</sup>. The number of iterations used was 2,000, plus a batch of 500 unretained iterations used to warm up the sampler. The results obtained include not only the posterior odds ratios (see Table 4.2), but also the posterior mean of the parameters, collected in Table 4.3, and the marginal posterior distributions which are graphed in Figures 4.1 to 4.5.

Table 4.2: Posterior odds ratios

	GDP	Cons'n.	Investm.	Employm.	M0
zero freq.	1.753	3.629	2.449	0.510	2.113
$\pi/2$ freq.	0.049	0.130	0.003	0.001	0.001
$\pi$ freq.	0.744	1.162	0.433	0.022	0.055

<sup>3</sup> All the Monte Carlo and Bayesian analysis computations were performed with software developed by the author and written in GAUSS 2.1. The preliminary analysis for the choice of the lag length was done by means of RATS 4.02 routines.

Table 4.3: Posterior means

	GDP	Cons'n	Investm.	Employm.	M0
$\alpha_0$	10.646	10.058	8.774	10.076	2.475
$\alpha_2$	-0.014	0.013	-0.002	0.001	0.002
$\alpha_1$	-0.014	0.027	0.045	0.001	-0.003
$\beta_1$	-0.019	-0.024	0.009	-0.003	0.001
$\gamma$	0.006	0.007	0.009	0.001	0.013
$\phi_1$	0.292	0.457	0.446	0.301	0.264
$\sigma$	0.022	0.016	0.036	0.005	0.008
$\psi_1$	-0.012	-0.005	-0.005	-0.008	-0.003
$\psi_2$	-0.148	-0.120	-0.160	-0.034	-0.618
$r$	0.175	0.151	0.186	0.032	0.309
$\theta$	-0.012	-0.354	-0.292	-0.988	-0.855

The results can be summarised as follows.

1) GDP: the posterior odds ratio analysis seems to clearly favour the hypothesis of zero frequency integration ( $POR=1.753$ ). The posterior mean of  $\psi_1$  is very low (-0.012), and its posterior distribution assigns high probability mass to the immediate neighbourhood of zero (see Figure 4.1). The posterior odds ratio instead soundly rejects the hypothesis of  $\pi/2$  frequency integration ( $POR=0.049$ ). This is confirmed by the value of the posterior mean of  $r$  (0.175) and by the shape of its posterior distribution, which assigns almost no weight to values near to zero. The possible presence of a  $\pi$  frequency unit root is more controversial ( $POR=0.74$ ). The posterior distribution of  $\psi_1$  assigns a non-negligible probability mass to values near to zero, although the mode of the distribution is well distant from zero. Considering all these results together, one might cautiously assume that the series

has a non-seasonal unit root, but that its seasonality might be dealt with by seasonal dummies. This contrasts with Osborn's results.

2) Consumption: the posterior odds ratio leads to clear acceptance of the long run unit root hypothesis ( $POR=3.629$ ). The posterior mean of  $\psi_1$  is close to zero (0.005), and the whole posterior distribution is concentrated near zero (see Figure 4.2). As for the  $\pi/2$  frequency integration hypothesis, ( $POR=0.134$ ), it is squarely rejected by the data, and the marginal posterior distribution of  $r$  gives all its weight to values well away from zero (posterior mean=0.151). Data are not conclusive on the issue of the presence of a  $\pi$  frequency unit root ( $POR=1.16$ ): the posterior distribution of  $\psi_2$  has mean equal to -0.120, mode equal to .006, but gives high weight to values near to zero. Varying the values of hyperparameters did not help to resolve uncertainty: the  $POR$  remained close to 1 for all the prior configurations being specified. Data are simply not very informative in this respect. Therefore one could weakly favour the presence of a bi-annual stochastic cycle in the data, but not the presence of an annual cycle. This again contrasts with Osborn's findings.

3) Investment: again for this series the presence of a zero frequency unit root seems unquestionable ( $POR=2.449$ ): the posterior distribution of  $\psi_1$  (see Figure 4.3) is squeezed to the immediate left neighbourhood of zero, with a posterior expectation of -0.012. The results contradicts the presence of a  $\pi/2$  frequency unit root, given that the  $POR$  is 0.003, and the posterior distribution of  $r$  does not assign any weight to the neighbourhood of zero; its posterior mean is 0.186. Likewise the model rejects the hypothesis of  $\pi$  frequency unit root ( $POR=0.433$ ), and the posterior distribution of  $\psi_2$  does not give the neighbourhood of zero substantial probability mass. For this series, one could thus conclude that, first, the series is  $I(1)$  in the conventional sense, and second, that non-stationary stochastic seasonality can be ruled out, receiving no support from the posterior analysis.



Deterministic seasonals account for the seasonal pattern. This is in perfect accordance with Osborn's results.

4) Employment: the presence of a zero frequency unit root is rejected by the data, the alternative hypothesis being preferred in the light of the *POR* (0.510). This is in sharp contrast with Osborn's findings concerning this series. The posterior distribution of  $\psi_1$  (see Figure 4.4) has mean -0.008, mode -0.009, and it does not give much weight to values near zero. Similarly, but more neatly, the posterior analysis reject the hypotheses of  $\pi$  and  $\pi/2$  frequencies integration (*POR* = 0.001 and 0.022 respectively). Also the posterior distributions of  $\psi_1$  and  $\psi_2$  are both clearly distant from zero. The series is therefore taken to be stationary around a deterministic linear trend with seasonal intercept shifts.

5) Real M0: the zero frequency integration hypothesis is clearly accepted on the basis of a *POR* of 2.113. On the contrary, both the hypotheses of seasonal integration are rejected on the basis of the posterior odds ratios (0.001 and 0.055, respectively). Also the examination of the posterior distributions of  $\psi_1$  and  $\psi_2$  are consistent with these findings (see Figure 4.5). This is consistent with Osborn's results.

As a partial corroboration of these findings, a sensitivity analysis experiment has been carried out. For the sake of brevity, I consider only the GDP series. Clearly, given the high dimensionality of the hyperparameter space, it is not feasible to monitor the effects of changes on all hyperparameters, and I focus only on the most crucial ones, that is, those controlling the prior distributions of  $\psi_1$ ,  $\psi_2$  and  $r$ . The prior hyperparameter specification (35), which produces the benchmark prior 1, is modified to generate another 4 priors along the following lines:

$$\text{prior 2: } \lambda_1 = \frac{2}{\tilde{\psi}_1}, \lambda_2 = \frac{2}{\tilde{\psi}_2}, \lambda_r = \frac{2}{\tilde{r}},$$

$$\text{prior 3: } \lambda_1 = \frac{4}{\tilde{\psi}_1}, \lambda_2 = \frac{4}{\tilde{\psi}_2}, \lambda_r = \frac{4}{\tilde{r}},$$

(36)

$$\text{prior 4: } \lambda_1 = \frac{0.5}{\tilde{\psi}_1}, \lambda_2 = \frac{0.5}{\tilde{\psi}_2}, \lambda_r = \frac{0.5}{\tilde{r}},$$

$$\text{prior 5: } \lambda_1 = \frac{0.25}{\tilde{\psi}_1}, \lambda_2 = \frac{0.25}{\tilde{\psi}_2}, \lambda_r = \frac{0.25}{\tilde{r}}.$$

Prior distributions 2 and 3 are more and more squeezed near zero values, whereas priors 4 and 5 assign greater weight to values distant from zero. For each one of these prior specifications, the posterior analysis described above was repeated out in exactly the same terms. The results in terms of the associated posterior odds ratios are presented in Table 4.4.

Table 4.4: Sensitivity analysis, GDP series

	prior 1	prior 2	prior 3	prior 4	prior 5
zero freq.	1.753	2.201	2.485	1.551	1.547
$\pi/2$ freq.	0.049	0.079	0.401	0.038	0.025
$\pi$ freq.	0.744	0.811	0.899	0.626	0.555

As one can easily see, these changes to the prior specification do not radically alter the nature of the results. As would be expected, priors 2 and 3 tend to give a higher posterior probability to the integration hypotheses, whereas priors 4 and 5 tend to favour the alternative hypotheses. These results seem encouraging and are interpreted as giving strength to the findings of this chapter.

Summing up, the Bayesian approach I propose is helpful in shedding new light on the inferential problem connected to the presence of unit roots at different frequencies. It is a sensible approach because it is based on a sensible parameterisation, and it allows a symmetric treatment of all the hypotheses being

tested. No use of asymptotics is made, and all the relevant posterior distributions are exact. In the particular application, the procedure seems to work well, giving in most cases a clear response to the issue of the presence of unit roots. The results seem to be robust with respect to alternative sensible prior specifications.

#### **[4.7] Conclusion**

The chapter presents a new testing procedure to ascertain the presence of unit roots at different frequencies in quarterly data. Given the weaknesses and logical inconsistencies of the classical inference setting, the proposed procedure is Bayesian, and relies on posterior odds ratio computations. Special emphasis is placed on devising a sensible prior distribution specification. The resulting joint posterior distribution is treated by means of a Gibbs sampling algorithm.

The procedure is applied to a set of UK series, previously analysed by Osborn (1990). In contrast to her results, less evidence was found in favour of non stationary stochastic seasonality, which seems to occur only for the consumption series. For the employment series it was found that the trend stationary alternative is preferred to the hypothesis of zero frequency integration: this series seem stationary around a deterministic time trend. All the other series are found  $I(1)$  in the traditional sense, that is they possess a zero frequency unit root, as in Osborn (1990).

#### **Appendix [4.A]: Proofs of distributional results**

##### **Proof of lemma 4.4.1**

The exponential term in (17) can be written as:

$$-\frac{1}{2\sigma^2}[(z-C\eta_1)'(z-C\eta_1) + (\beta^*-b^*)' V^{*-1}(\beta^*-b^*)],$$

where  $z$  is a  $(T \times 1)$  vector with  $t$ th element:

$$\phi^*(L)z_{4t} - \psi_1 z_{1t-1} - \psi_2 z_{2t-1} - 2r[(\sin \theta) z_{3t-1} - (\cos \theta) z_{3t-2}],$$

$C$  is a  $(T \times 5)$  matrix with  $t$ th row:

$$[-4\psi_1 - 4\psi_2 \cos(\pi t), 4r \cos[(\pi/2)t - \theta], 4r \sin[(\pi/2)t - \theta], 4\phi^*(1) + 10\psi_1 + 2\psi_2 + 4r(\sin \theta - \cos \theta) - 4\psi_1 t].$$

Defining:

$$z^* = [V^{*-1/2} b^*, z]', \quad C^* = [V^{*-1/2}, C]',$$

the exponential term in the joint posterior can be written as:

$$-\frac{1}{2\sigma^2}[(z^*-C^*\eta_1)'(z^*-C^*\eta_1)].$$

Therefore I have:

$$p(\eta_1 | \bar{\eta}_1, D_T) \sim N[(C^*C^*)^{-1}C^{*'}z^*, \sigma^2(C^*C^*)^{-1}],$$

i.e. a 5-variate normal distribution, whose position vector and dispersion matrix depend on the other parameters of the system.

#### Proof of lemma 4.4.2

Consider the joint posterior pdf (17). It is evident that this depends on  $\eta_2$  only through the term  $e'e$  in the exponential part, which can be written as:

$$e'e = (y - X\eta_2)'(y - X\eta_2),$$

where  $y$  is  $(T \times 1)$  with  $i^{\text{th}}$  element given by:

$$y_{4i} = \psi_1 y_{1i-1} - \psi_2 y_{2i-1} - 2r[(\sin \theta)y_{3i-1} - (\cos \theta)y_{3i-2}],$$

and  $X$  is a  $(T \times k^*)$  matrix with  $k^*$  lags of  $y_{4i}$  in its  $i^{\text{th}}$  row.

Thus, conditionally on the other parameters of the system,  $\eta_2$  has the following posterior pdf:

$$p(\eta_2 | \bar{\eta}_1, D_T) \sim N[(X'X)^{-1}X'y, \sigma^2(X'X)^{-1}].$$

#### Proof of lemma 4.4.3

From expression (17) I have that :

$$p(\eta_2 | \bar{\eta}_1, D_T) \propto \sigma^{-T-5} \exp \{ -c/(2\sigma^2) \}, \quad c = e'e + (\beta^* - b^*)' V^{-1} (\beta^* - b^*),$$

where  $c$  depends on the data and on  $\bar{\eta}_1$ . Given that the expression above is an inverted-Gamma distribution (see Zellner, 1971, p.371):

$$p(y | \nu, \alpha) = 2[\Gamma(\alpha)\gamma^\alpha y^{2\alpha-1}]^{-1} \exp\{-1/(2\gamma y^2)\},$$

the connections between inverse Gamma, Gamma and  $\chi^2$  distributions can be exploited. We define  $\sigma' = c/\sigma^2$ , and since  $\sigma'$  is a monotone function of  $\sigma$ , we conclude that:

$$p(\sigma' | \bar{\eta}_3, \mathbf{D}_T) \propto \sigma'^{(T+2)/2} \exp\{-\sigma'/2\},$$

i.e. that the conditional posterior pdf of  $\sigma'$  given all the other parameters is  $\chi^2(T+4)$ . This is intuitive, since:

$$c/\sigma^2 = (\mathbf{e}'\mathbf{e} + \sigma_\theta^2(\theta - \mu_\theta)^2)/\sigma^2 = \sigma^{-2} \left[ \sum_{i=1}^T e_i^2 + (\boldsymbol{\beta} - \mathbf{b})' \mathbf{V}^{-1}(\boldsymbol{\beta} - \mathbf{b}) + \sigma_\theta^2(\theta - \mu_\theta)^2 \right],$$

is just the sum of the squares of  $T+4$  independent standard normal variates.

#### Proof of lemma 4.4.4

Starting from expression (17), it is easy to see that the parameter  $\psi_1$  appears both in the exponential term, via  $\mathbf{e}'\mathbf{e}$ ,  $(\alpha_0 - a_0)^2 \psi_1$ , and  $\lambda_1 \psi_1$ , and outside, via  $(-\psi_1)^{-1/2}$ . The term  $\mathbf{e}'\mathbf{e}$  in the exponential part can be represented as:

$$\mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{x}\boldsymbol{\eta}_4)' (\mathbf{y} - \mathbf{x}\boldsymbol{\eta}_4),$$

where the vector  $(T \times 1)$   $\mathbf{y}$  has  $i^{\text{th}}$  element equal to  $\phi^*(L)y_{4i} - y_{2i-1} - 2r[(\sin \theta)y_{3i-1} - (\cos \theta)y_{3i-2}]$ , and  $\mathbf{x}$  is another  $(T \times 1)$  vector with corresponding element equal to  $y_{1i-1}$ .

On the basis of this representation, and using a notation consistent with expression above, I can represent the whole relevant exponential term as quadratic in  $\boldsymbol{\eta}_4$ :

$$p(\eta_4 | \bar{\eta}_4, \mathbf{D}_T) \propto (-\eta_4)^{1/2} \exp\left\{-\frac{1}{2\tau_1^2} (\eta_4 - \mu_1)^2\right\},$$

$$\mu_1 = [\mathbf{x}'\mathbf{y} + (1/2)(\alpha_0 - a_0)^2 + \lambda_1 \sigma^2] / (\mathbf{x}'\mathbf{x}), \quad \tau_1 = \sigma^2 / (\mathbf{x}'\mathbf{x}).$$

Although Gaussian-looking in the exponential part, the above distribution is not unfortunately of any analytically known form.

As for  $\eta_5$  (i.e.  $\psi_2$ ), I have:

$$p(\eta_5 | \bar{\eta}_5, \mathbf{D}_T) \propto (-\eta_5)^{1/2} \exp\left\{-\frac{1}{2\tau_2^2} (\eta_5 - \mu_2)^2\right\},$$

$$\mu_2 = [\mathbf{x}'\mathbf{y} + (1/2)(\alpha_2 - a_2)^2 + \lambda_2 \sigma^2] / (\mathbf{x}'\mathbf{x}), \quad \tau_2 = \sigma^2 / (\mathbf{x}'\mathbf{x}).$$

Here, the vectors  $\mathbf{y}$  and  $\mathbf{x}$  have been conveniently defined to decompose:

$$\mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{x} \eta_5)' (\mathbf{y} - \mathbf{x} \eta_5).$$

Hence, for the parameter  $\eta_5$  I have results that coincide with those seen for  $\eta_4$  ( $\psi_1$ ). The parameters  $\tau_2$  and  $\mu_2$  derive from a similar sort of decomposition of the exponential part as seen above.

The conditional posterior pdf of  $\eta_6$  (that is  $r$ ) is likewise complicated:

$$p(\eta_6 | \bar{\eta}_6, \mathbf{D}_T) \propto \eta_6 \exp\left\{-\frac{1}{2\tau_r^2} (\eta_6 - \mu_r)^2\right\},$$

$$\mu_r = \{\mathbf{x}'\mathbf{y} - [(\alpha_1 - a_1)^2 + (\beta_1 - b_1)^2] / 2 + \lambda_r \sigma^2\} / (\mathbf{x}'\mathbf{x}), \quad \tau_r = \sigma^2 / (\mathbf{x}'\mathbf{x}),$$

$$\mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{x} \eta_6)' (\mathbf{y} - \mathbf{x} \eta_6).$$

The conditional posterior of  $\eta_7$  ( $\theta$ , the phase angle) is even more complicated:

$$p(\eta_7 | \bar{\eta}_7, \mathbf{D}_T) \propto \exp\left\{\left(-\frac{1}{2\sigma^2} [\mathbf{e}'\mathbf{e} + (\sigma/\sigma_\theta)^2 (\theta - \mu_\theta)^2]\right)\right\},$$

$$e'e = \sum_{t=1}^T [\phi'(L) y_{4t} - \psi_1 y_{1t-1} - \psi_2 y_{2t-1} - (\sin \theta)(2r y_{3t-1}) + (\cos \theta)(2r y_{3t-2})]^2.$$

We obtain draws from these conditional posterior distribution using the algorithms described in Appendix [4.B].

#### **Appendix [4.B] : Rejection Sampling from the Conditional Posterior Distributions**

I follow the approach of Geweke (1994). A brief description of the method used is given in Chapter 3. In this appendix, the solutions adopted to the particular problem being treated are developed with particular attention to their capability of providing efficient random drawings from the conditional posterior pdfs.

Given a non standard pdf to draw from,  $f(x|\theta)$ , the problem is that of optimally choosing the comparison (or 'envelope') function  $g(x|\phi)$ . The aim is to maximise computational efficiency, i.e. to maximise the unconditional probability of retaining draws from the comparison function. We have thus to solve:

$$\min_{\phi} \left[ \max_x \left( \frac{f(x|\theta)}{g(x|\phi)} \right) \right].$$

We emphasise that the choice of the parameters in  $\phi$ , together with the choice of the functional form of  $g(\cdot)$  does not affect the correctness of the results from the synthetic draws, but only affects the efficiency of the procedure, i.e. the rejection rate of the draws from  $g(\cdot)$ , and hence the computational time<sup>4</sup>. In the remainder of the present appendix, the choices made in this respect are discussed for each of

<sup>4</sup> The computations necessary to the analysis of each series took approximately two hours on a 386 20 MHz PC.



the 4 synthetically replicated conditional posterior pdf's, namely those of  $\psi_1$ ,  $\psi_2$ ,  $r$  and  $\theta$ .

1) Conditional posterior density of  $\psi_1$ . The distribution is:

$$f(x) \propto (-x)^{1/2} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}, x \in \mathbb{R}_-,$$

with the quantities  $\mu$  and  $\sigma$  defined as in Section 4.4. The comparison function chosen is:

$$g(x) \propto \exp\left\{-\frac{1}{2\sigma^2}(x-\nu)^2\right\} I_{(-\infty,0)}(x),$$

a negative truncated normal distribution. In order to avoid further complications, I confine to the choice to the location parameter  $\nu$ .

The differentiation of  $\log f() - \log g()$  with respect to  $x$  gives:

$$x^* = -\sigma^2/[2(\mu-\nu)].$$

Provided  $\mu - \nu > 0$ ,  $x^*$  belongs to the support of  $f(x)$ . The second order condition for a maximum holds.

The expression  $\log g() - \log f()$  evaluated in  $x^*$  is maximised with respect to  $\nu$ , yielding:

$$\nu = (\mu - (\mu^2 + 2\sigma^2)^{1/2})/2.$$

This is the only admissible solution. The second order condition is satisfied.

2) Conditional posterior pdf of  $\psi_2$ . Exactly the same computations as above apply.

3) Conditional posterior pdf of  $r$ . The distribution is:

$$f(x) \propto x \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}, x \in \mathbb{R}_+,$$

with the quantities  $\mu$  and  $\sigma$  defined as in Section 4.4. The comparison function chosen is:

$$g(x) \propto \exp\left\{-\frac{1}{2\sigma^2}(x-\nu)^2\right\} I_{(0,+\infty)}(x),$$

a positive truncated normal distribution. The same kind of computations as previously described yield:

$$x^* = \sigma^2/(\nu - \mu), \nu = [\mu + (\mu^2 + 4\sigma^2)^{1/2}]/2.$$

4) Posterior distribution of  $\theta$ . Recalling the analysis contained in Section 4, I can write:

$$\begin{aligned} f(\theta) &\propto \exp\left\{-\frac{1}{2\sigma^2}[\mathbf{e}'\mathbf{e} + (\sigma/\sigma_\theta)^2(\theta - \mu_\theta)^2]\right\}, \theta \in [-\pi, \pi], \\ \mathbf{e}'\mathbf{e} &= \sum_{t=1}^T [\phi^*(L)y_{4t} - \psi_1 y_{1t-1} - \psi_2 y_{2t-1} - (\sin \theta)(2r y_{3t-1}) + (\cos \theta)(2r y_{3t-2})]^2 \\ &= (\mathbf{y} - \mathbf{X}\boldsymbol{\omega})(\mathbf{y} - \mathbf{X}\boldsymbol{\omega}), \boldsymbol{\omega} = [\sin \theta, \cos \theta]', \mathbf{X} = \{\mathbf{x}_t'\}_{t=1}^T, \mathbf{x}_t' = [2r y_{3t-1} | -2r y_{3t-2}] \end{aligned}$$

The problem of obtaining draws looks quite cumbersome, given that the maximisation of  $\log f()$  -  $\log g()$  involves trigonometric expressions. An easy way out is to choose:

$$g(\theta) \propto \exp\left\{-\frac{1}{2\sigma_\theta^2}(\theta - \mu_\theta)^2\right\} I_{(-\pi, +\pi)}(\theta),$$

so that the function to be maximised is:

$$\log f() - \log g() = -1/(2\sigma^2) (\mathbf{y} - \mathbf{X}\boldsymbol{\omega})(\mathbf{y} - \mathbf{X}\boldsymbol{\omega}).$$

We maximise this expression with respect to  $\boldsymbol{\omega}$ , under the non-linear constraint:  $\boldsymbol{\omega}'\boldsymbol{\omega} = 1$ . This can be done numerically. Once the maximum value is obtained, say  $m$  one then draws  $y$  from  $U(0, m)$ , and applies the rejection sampling technique.

#### Appendix [4.C] : Proofs of the Smooth Transition Results

Starting from the comparison between the hypotheses  $H_A: \psi_1 = 0$  vs.  $H_B: \psi_1 < 0$ , it is easy to see that the joint posterior pdf under  $H_B$  can be marginalised with respect to  $\alpha_0$ , yielding:

$$\begin{aligned} p(\bar{\alpha}_0 | \mathbf{D}_T, H_B) &\propto \sigma^{-(T+4)} (-\psi_2)^{1/2} r [1 - 16 T \psi_1]^{-1/2} \\ &\exp\left\{-\frac{1}{2\sigma^2} [\mathbf{w}^* \mathbf{M}(\mathbf{X}^*) \mathbf{w}^* + (\boldsymbol{\beta}^{**} - \mathbf{b}^{**})' \mathbf{V}^{*-1} (\boldsymbol{\beta}^{**} - \mathbf{b}^{**}) \right. \\ &\quad \left. + \sigma_\theta^2 (\theta - \mu_\theta)^2] + \lambda_2 \psi_2 - \lambda, r\right\} \end{aligned} \quad (37)$$

where  $\boldsymbol{\beta}^{**} = [\alpha_2, \alpha_1, \beta_1]'$ ,  $\mathbf{b}^{**} = [a_2, a_1, b_1]'$ ,  $\mathbf{V}^{**} = \text{diag}(-\psi_2^{-1}, r^{-1}, r^{-1}, (\sigma_y/\sigma)^2)$ , and the  $[(T+1) \times 1]$  vectors  $\mathbf{w}^*$  and  $\mathbf{X}^*$  defined as:

$$\begin{aligned} \mathbf{w}^* &= [-\psi_1^{1/2} \alpha_0, \mathbf{w}']', \quad \mathbf{X}^{*'} = [-\psi_1^{1/2}, \mathbf{X}']', \quad \mathbf{w} = \{\mathbf{w}_t\}_{t=1}^T, \quad \mathbf{X} = \{\mathbf{x}_t\}_{t=1}^T, \\ \mathbf{w}_t &= \phi^*(L) y_{4t} - \psi_1 y_{1t-1} - \psi_2 y_{2t-1} - 2r [(\sin \theta)(z_{3t-1} + 2\gamma) - (\cos \theta)(z_{3t-2} + 2\gamma)], \\ \mathbf{x}_t &= -4 \psi_1. \end{aligned}$$

It is therefore easy to see that

$$\lim_{r \rightarrow 0} [1 - 16 T \psi_1]^{-1} = 1, \quad \lim_{r \rightarrow 0} [\mathbf{w}^* \mathbf{M}(\mathbf{X}^*) \mathbf{w}^*] = \mathbf{e}_A' \mathbf{e}_A,$$

where  $\mathbf{e}_A$  is the vector of error terms under  $H_A$ , i.e.  $e_{At} = \phi^*(L)y_{4t} - \psi_2 y_{2t-1} - 2r[(\sin \theta)(z_{3t-1} + 2\gamma) - (\cos \theta)(z_{3t-2} + 2\gamma)]$ . This establishes the first smooth transition result.

Comparing the hypotheses  $H_A: \psi_2 = 0$  vs.  $H_B: \psi_2 < 0$ , a very similar strategy is used to prove smooth transition. It is necessary to preliminarily marginalise the joint posterior pdf under  $H_B$  with respect to  $\alpha_2$ . This can be done analytically

Finally, comparing the hypotheses  $H_A: r = 0$  vs.  $H_B: r > 0$ , it is necessary to analytically marginalise the joint posterior pdf with respect to  $\alpha_1$  and  $\beta_1$ . When  $r = 0$ , it is very easy to marginalise the resulting posterior pdf with respect to  $\theta$ , obtaining the required smooth transition result.

Figure 4.1: UK GDP results

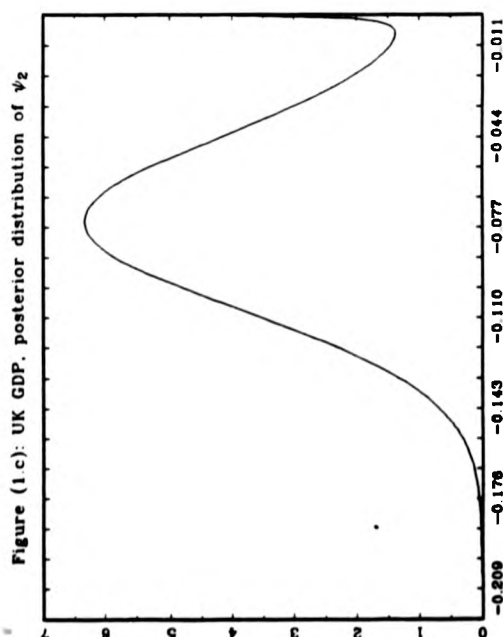
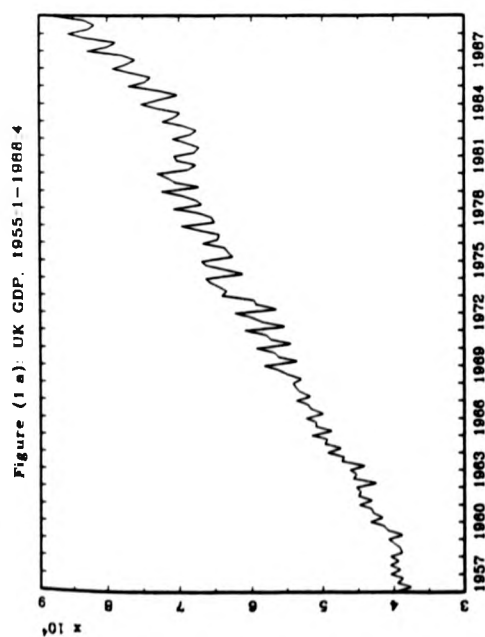
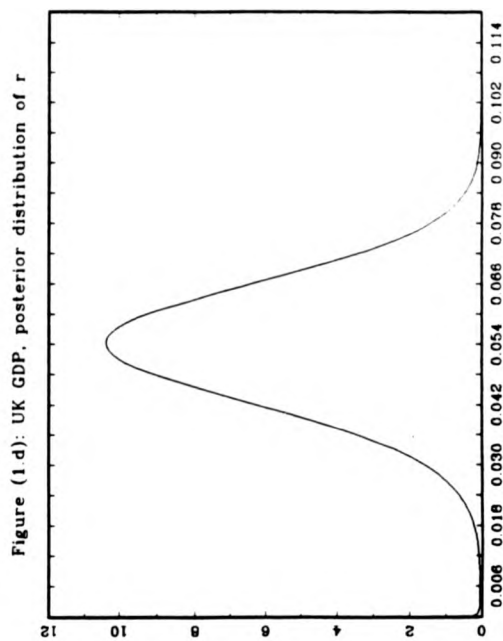
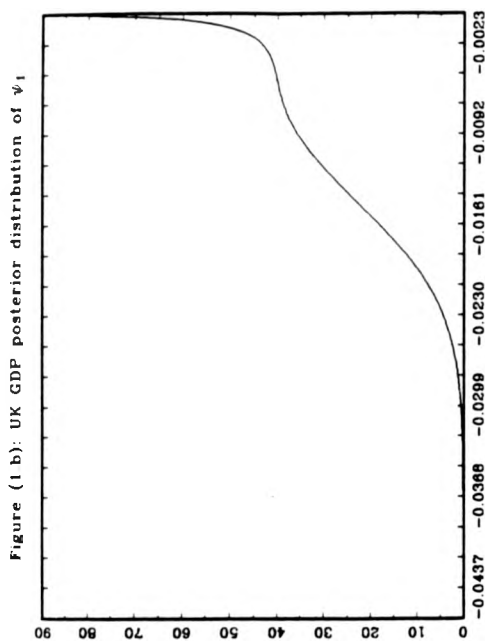


Figure 4.2: UK consumption results

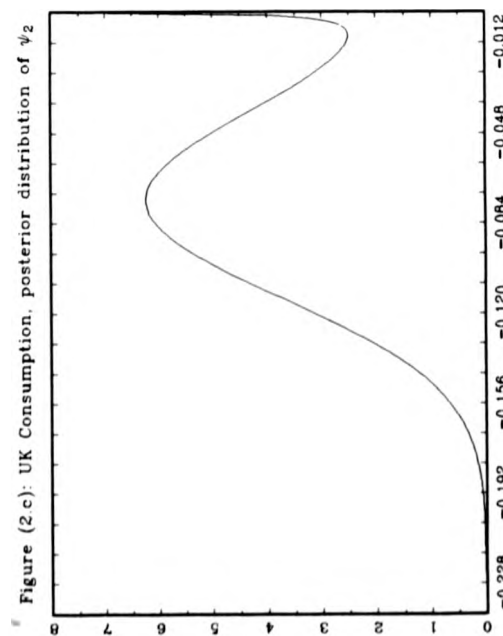
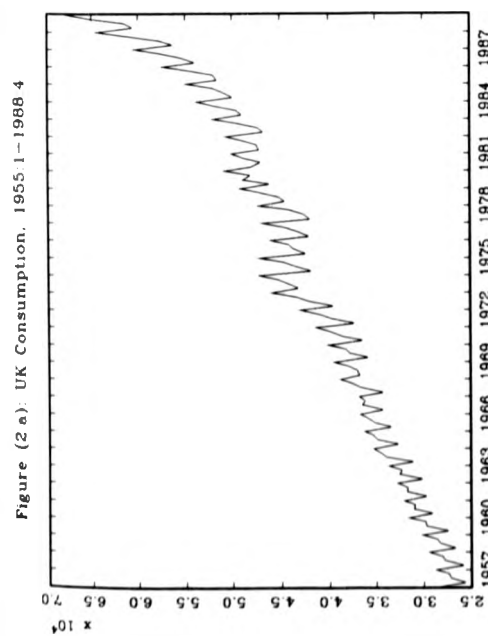
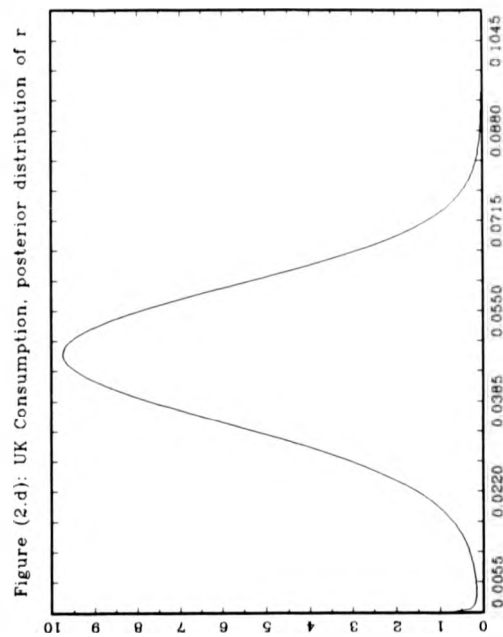
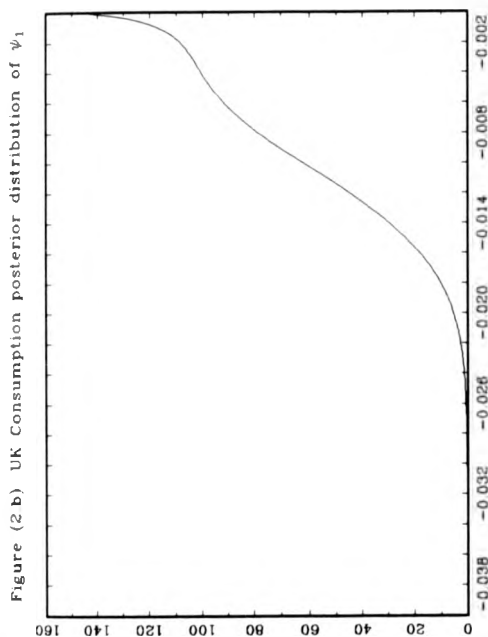


Figure 4.3: UK investment results

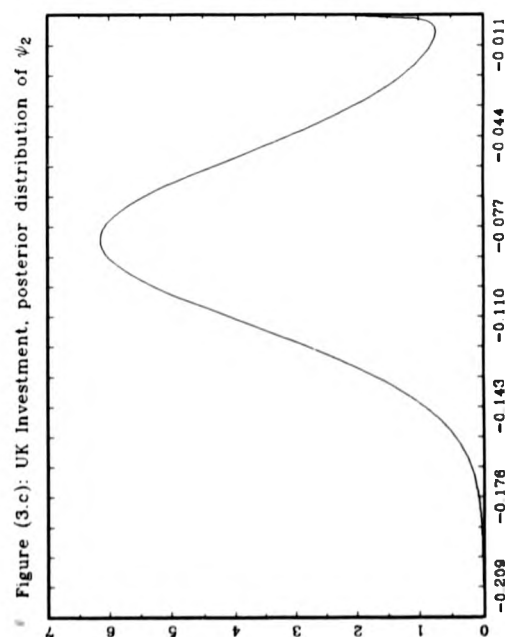
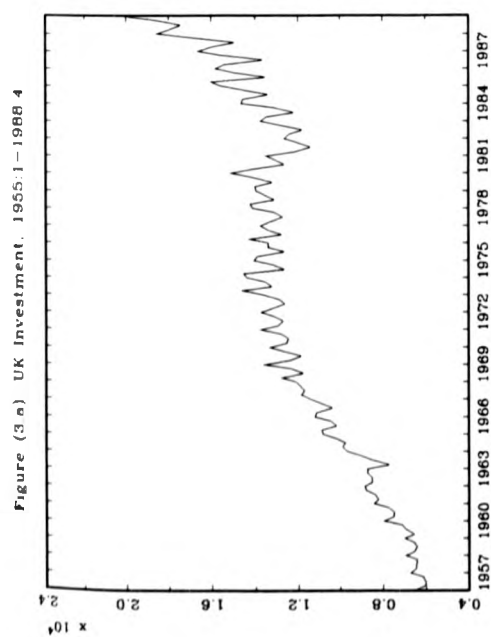
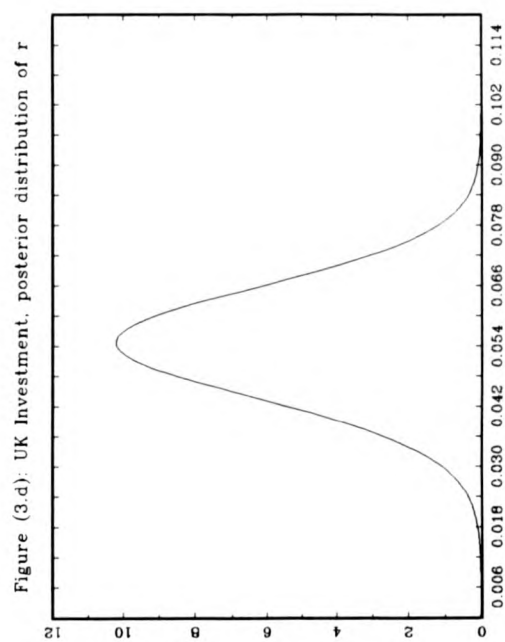
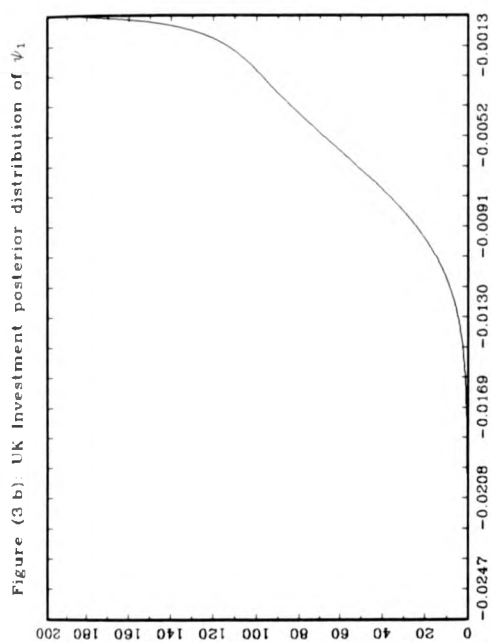


Figure 4.4: UK employment results

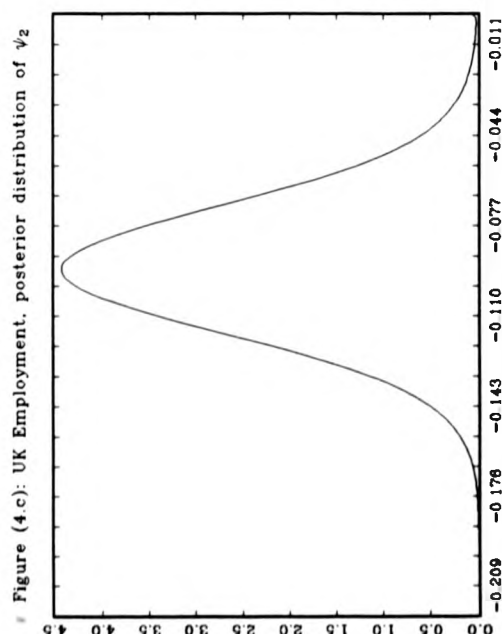
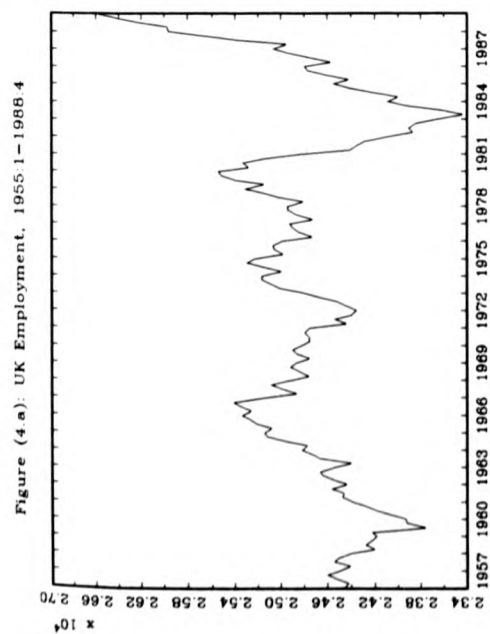
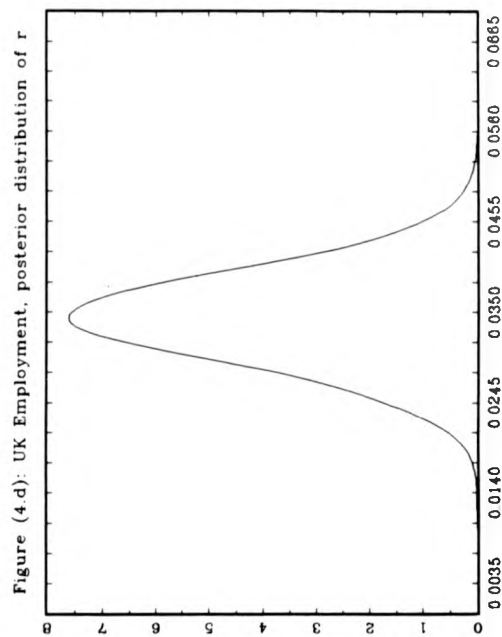
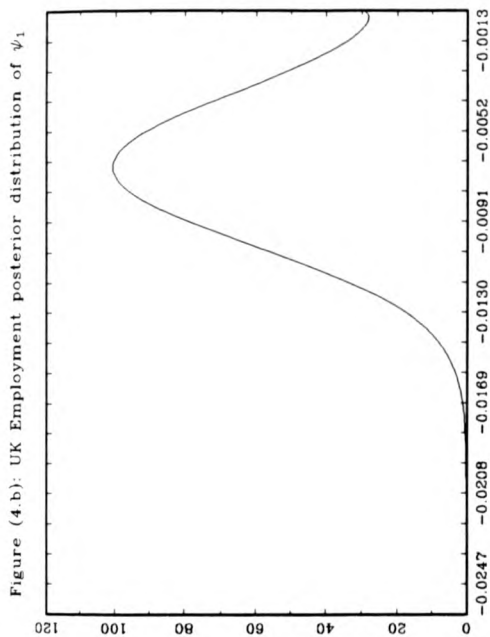
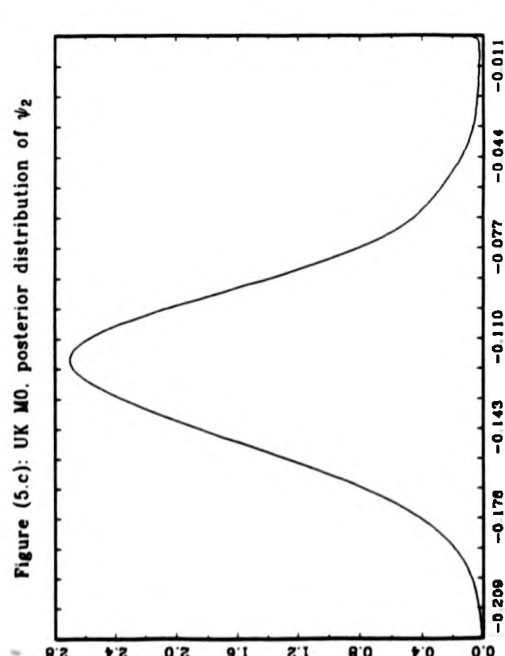
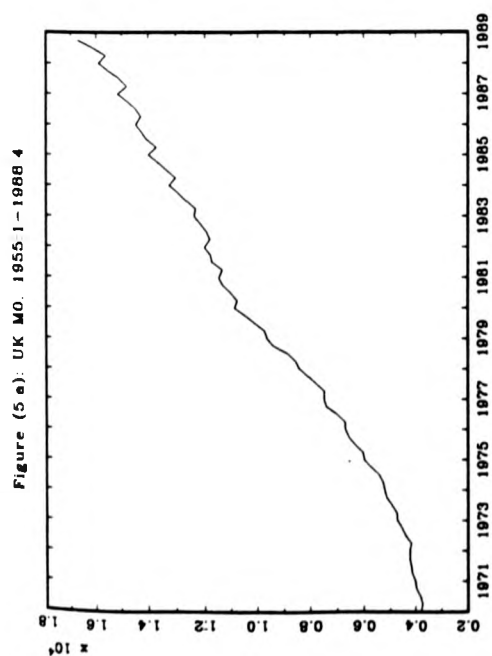
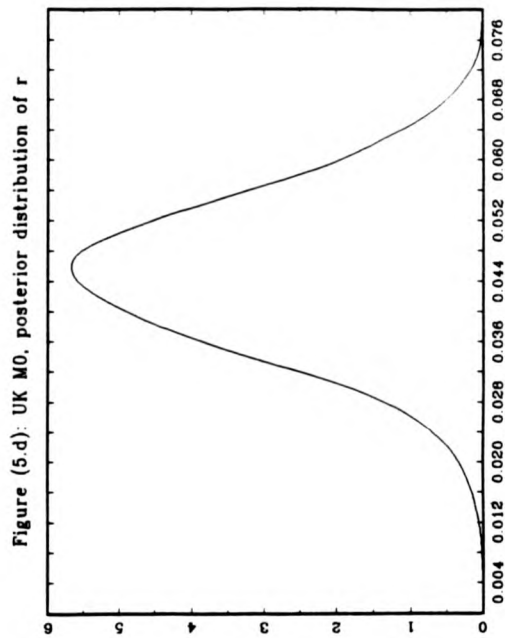
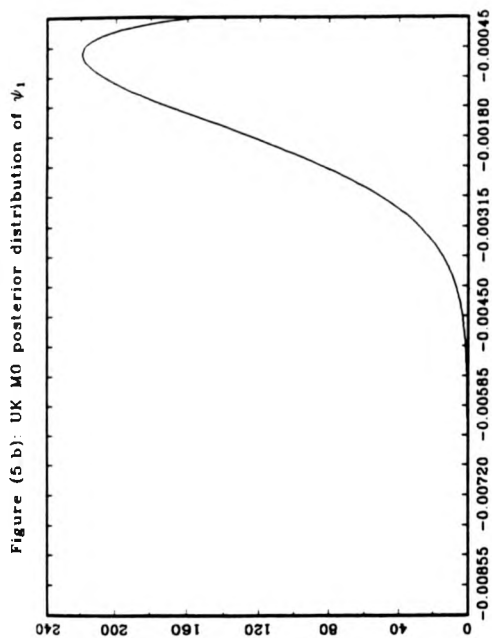




Figure 4.5: UK M0 results



## **PART II: The Multivariate Analysis of Non Stationary Time Series**

### **Chapter 5: Non Stationarity in Multivariate Time Series Analysis. The Classical Approach.**

#### **[5.0] An Overview of the Chapter.**

In the present chapter I discuss the problems encountered in the analysis of the interactions between non-stationary series. In the first section of the chapter, the notion of cointegration is discussed in the light of its contrast with the concept of "spurious regression". The main representations of cointegrated systems are briefly described in Section [5.2], in order to understand fully the different properties of a cointegrating system. Section [5.3] deals with the main estimation techniques available to estimate cointegrating relationships, with particular attention being devoted to the maximum likelihood analysis put forward by S. Johansen, since this is the only approach capable of delivering a testing procedure in order to test for the number of long-run relationships. Section [5.4] is devoted to the controversial issue of the interpretation of the estimated cointegrating relationships, discussing the relevant identification conditions and the possibility of testing the validity of the over-identifying constraints. Section [5.5] reviews the available asymptotic results which are the basis of the inferences being drawn in applied studies, and the last section of the chapter discusses the corresponding finite sample distributional results obtained via analytical and simulation studies. In my view, the sharp contrast between the asymptotic and finite sample properties of the estimators and testing procedures provides one of the main motivations for the use of inferential techniques based on finite sample properties. A Bayesian approach allows one to

do so, being crucially based on the posterior finite sample distributions of functions of interest of the parameters.

### [5.1] Spurious Regression and Cointegration

The issue of the interpretation of results of regressions among non-stationary variables goes back to the discussion of "nonsense regression" by Yule, (1926), and the famous contribution by Granger and Newbold (1974), who refer instead to "spurious regression". The notion of spurious regression relates to a regression among non stationary variables, when good measures of fit may be found even in the absence of any direct links among the variables. This was shown with Monte Carlo simulations by Granger and Newbold (1974), and proved analytically by Phillips (1986). A very simple example of spurious regression can be provided by considering two unrelated univariate random walk processes:

$$\Delta y_{1t} = \varepsilon_{1t}, \Delta y_{2t} = \varepsilon_{2t}, \text{ with } E(\varepsilon_{it} \varepsilon_{jt}) = 0, \forall i \neq j, s \neq t. \quad (1)$$

The regression:

$$y_{1t} = \beta_0 + \beta_1 y_{2t} + e_t \quad (2)$$

would yield an  $R^2$  index asymptotically different from zero and all the tests on the parameters (the  $t$ -tests on  $\beta_0$ ,  $\beta_1$  and the joint  $F$ - tests) would have diverging limiting distributions with asymptotic size equal to one. This circumstance would clearly lead to wrong inferential conclusions being drawn on the basis of any sample, no matter how large. Hence the suggestion of Granger and Newbold was

to difference all variables prior to the analysis in order to eliminate the occurrence of the problem just described. Of course, this would preclude the possibility of obtaining any information on the long run relationships among the non stationary variables being analysed.

Long run relationships themselves are particularly interesting because they immediately relate to the notion of equilibrium links among sets of economic variables. By equilibrium is meant a state from which there is no endogenous tendency to deviate. The concept of cointegration was formalised by Granger (1981) and Engle and Granger (1987), and refers to a statistical feature of non stationary series which easily lends itself to meaningful interpretations in terms of the existence of such equilibrium relationships.

In its simplest formulation, the definition of cointegration is as follows: given  $y_t$ , a  $(n \times 1)$  vector of  $I(d)$  variables, they are said to be *cointegrated* with orders  $(d, b)$  and with rank  $r < n$  if there exist a full rank  $(n \times r)$   $\beta$  matrix such that  $z_t = \beta'y_t$  is  $I(d-b)$ . This means that there exist  $r$  linear combinations of the elements of  $y_t$  which generate variables with a lower order of integration.

The case most intensely studied in the literature is when  $d=b=1$ , i.e. when  $y_t$  is  $I(1)$  and the  $z_t$  variables are stationary. In this circumstance, it is immediate to consider the columns of  $\beta$  as the weights of different equilibrium relationships, and the elements of  $z_t$  as the disequilibrium errors. Equilibrium relationships are relevant only if disequilibrium errors are stationary, i.e. if they are mean-reverting or, in other words, shocks that make variables deviate from their equilibrium relationships are not persistent.

To give a very simple example of this, consider two  $I(1)$  variables,  $x_{1t}$  and  $x_{2t}$ , and imagine that there exists a linear long run equilibrium relationship between them of the kind:  $x_{1t}^* = \beta_2 x_{2t}^*$ .

If the equilibrium relationship is relevant in determining the joint behaviour

of  $x_{1t}$  and  $x_{2t}$ , the disequilibrium errors should be stationary, i.e. the series  $z_t = \beta'x_t = [1, -\beta_2][x_{1t}, x_{2t}]'$  should be stationary. This would imply  $x_{1t}$  and  $x_{2t}$  being cointegrated with rank equal to one.

On the other hand, a regression among  $I(1)$  variables in the absence of equilibrium relationships would be associated with non-stationary disturbances. This circumstance is then the hallmark of spurious regressions. In fact, taking for example the *DGP* (1), it is immediate to realise that:

$$e_t = \sum_{j=1}^t \varepsilon_{1j} - \beta_0 - \beta_1 \sum_{j=1}^t \varepsilon_{2j},$$

which is clearly a non stationary process.

## [5.2] Representation and identification issues

In this section I will review the main representation results concerning cointegrated  $I(1)$  variables, directly drawing from the Granger representation theorem, as stated in Engle and Granger (1987). Let us consider a  $n$ -dimensional  $k^{\text{th}}$  order VAR process of the kind:

$$A(L) y_t = \mu_0 + \varepsilon_t, \quad A(L) = I_n - A_1 L - A_2 L^2 - \dots - A_k L^k, \quad (3)$$

$$E(\varepsilon_t) = 0 \quad \forall t, \quad E(\varepsilon_t \varepsilon_s') = \Sigma \quad \forall t, \quad E(\varepsilon_t \varepsilon_s') = 0 \quad \forall t \neq s.$$

In this model the deterministic part has been kept deliberately simple for exposition purposes. Below, I will treat the issue of different, more fully articulated, deterministic components.

Suppose that the following conditions are fulfilled:

(i)  $|A(L)| = 0$  has either unit roots or roots greater than one in modulus. This condition ensures that the non stationarity of the data can be removed by differencing. The matrix autoregressive polynomial has  $nk$  roots; some of them are unity and the remaining ones are stationary.

(ii) The matrix  $A(1)$  has rank equal to  $r < n$ . This means that it can be written as the product of two full rank  $(n \times r)$  matrices  $\alpha$  and  $\beta$ :

$$A(1) = -\alpha \beta'$$

This condition reflects the presence of  $r$  cointegration relationships. It ensures that the number of unit roots in the system is equal to  $s = n - r$ .

(iii) The  $(s \times s)$  matrix  $\alpha_{\perp}' \Psi \beta_{\perp}$  has full rank  $s$ , where  $\alpha_{\perp}' \alpha = \beta_{\perp}' \beta = 0$  and  $\Psi = -\left[ \frac{\partial A(z)}{\partial z} \right]_{z=1}$  i.e. the mean-lag matrix of the VAR representation. This condition

rules out the occurrence that some of the elements of  $y_t$  could be  $I(2)$  processes.

Under these conditions, the following results can be proved (see for example Banerjee *et al.*, 1993, and Johansen 1995a):

1)  $\Delta y_t$  and  $z_t = \beta' y_t$  are  $I(0)$  processes.

2) The expected values of these stationary processes are respectively:

$$E(\Delta y_t) = \beta_{\perp} (\alpha_{\perp}' \Psi \beta_{\perp})^{-1} \alpha_{\perp}' \mu_0, \text{ and}$$

$$E(\beta' y_t) = -(\alpha' \alpha)^{-1} \alpha' \mu_0 + (\alpha' \alpha)^{-1} (\alpha' \Psi \beta_{\perp}) (\alpha_{\perp}' \Psi \beta_{\perp})^{-1} \alpha_{\perp}' \mu_0.$$

3) The VAR system can be cast in an isomorphic error correction form:

$$\Gamma(L) \Delta y_t = \mu + \alpha \beta' y_{t-1} + \varepsilon_t \quad (4)$$

$$\Gamma(L) = I_p - \Gamma_1 L - \Gamma_2 L^2 - \dots - \Gamma_{k-1} L^{k-1}, \Gamma_1 = - \sum_{j=1}^k A_j.$$

4) There exists a moving average representation:

$$\Delta y_t = C(L) (\mu_0 + \varepsilon_t), \quad (5)$$

where:

$$C(1) = \beta_{\perp}(\alpha_{\perp}'\psi\beta_{\perp})^{-1}\alpha_{\perp}'. \quad (6)$$

5) It is possible to obtain a multivariate Beveridge-Nelson (Beveridge and Nelson, 1981) decomposition:

$$y_t = y_0 + G\xi_t + \tau t + C^*(L)\varepsilon_t, \quad (7)$$

$$G = \beta_{\perp}(\alpha_{\perp}'\psi\beta_{\perp})^{-1}, \quad \xi_t = \alpha_{\perp}' \sum_{j=1}^t \varepsilon_j, \quad \tau = C(1)\mu_0, \quad \Delta C^*(L) = C(L) - C(1).$$

Some brief comments on these results seem necessary in order to fully understand their implications. First of all, result (1) establishes that the processes  $\Delta y_t$  and  $\beta'y_t$  are stationary while  $y_t$  is not, and result (2) gives the analytical expression of the unconditional expected values of these two stationary vector processes.

Result (3) allows one to write the *VAR* representation in an equivalent form which is particularly useful for estimation purposes. This representation is nevertheless affected by lack of identification. In fact, by choosing any invertible  $(r \times r)$  matrix  $Q$ , it is possible to write:

$$\Gamma(L)\Delta y_t = \mu + \alpha^* \beta^* y_{t-1} + \varepsilon_t$$

where  $\alpha^* = \alpha Q^{-1}$  and  $\beta^* = \beta Q'$ . In order to identify  $\alpha$  and  $\beta$ , it is necessary to choose a normalisation, i.e. a unique choice of the matrix  $Q$ . A widely used normalisation consists in conceptually choosing  $Q = \beta_1^{-1}$  where  $\beta_1$  is the upper  $(r \times r)$  block of  $\beta$ . In this way the normalised  $\beta^*$  matrix is  $\beta^* = [I_r \mid \beta_2'']'$ , where  $\beta_2'' = \beta_2 \beta_1^{-1}$ . The result of this normalisation is sometimes referred to as Phillips' triangular representation, after Phillips (1991b). On the significance of this identification problem, I will return in section [5.3].

Result (4) gives the impulse response function of a cointegrated system. Notice that the long run impulse response coefficients are given by the matrix  $C(1)$  which

has reduced rank equal to  $s$ . Moreover, on the basis of (6), it is easy to see that  $\beta'C(1) = 0$ ; therefore, the effects of shocks on  $z_t$  die away as time elapses.

Results (5) is particularly important in order to understand the statistical properties of the cointegrated system. Notice that the system is driven by a  $s$ -dimensional random walk process  $\xi_t$ . The elements of this process are the "common stochastic trends" (Stock and Watson, 1988) determining the non-stationary behaviour of  $y_t$ . Choosing  $A = (\beta, \tau, \gamma)$  as a basis of  $R_n$ , with  $\gamma$  orthogonal to  $\beta$  and  $\tau$ , it is easy to see that along the directions of the subspace spanned by the columns of  $\gamma$  the process  $y_t$  behaves as a  $s-1$ -dimensional driftless random walk, whereas along the directions given by the columns of  $\beta$   $y_t$  is a stationary process without any deterministic trend.

In the description of the properties of a cointegrated system, I have chosen to start from model (3), which is clearly suitable for linearly trending variables. Clearly, different alternative specifications for the deterministic part are possible, accounting for different deterministic properties of the series being modelled. Following Johansen and Juselius (1990), it is possible to start from a cointegrated VAR model with a linear trend:

$$A(L) y_t = \mu_0 + \mu_1 t + \varepsilon_t, \quad A(L) = I_p - A_1 L - A_2 L^2 - \dots - A_k L^k, \quad \varepsilon_t \sim VWN(0, \Sigma).$$

In this case, the moving average representation is:

$$\Delta y_t = C(L) (\mu_0 + \mu_1 t + \varepsilon_t),$$

and the Beveridge-Nelson representation is:

$$y_t = y_0 + G\xi_t + \tau_0 t + \tau_1 t^2 + C^*(L) \varepsilon_t, \\ G = \beta_\perp (\alpha_\perp' \psi \beta_\perp)^{-1}, \quad \Delta C^*(L) = C(L) - C(1),$$



$$\xi_t = \alpha_{\perp}' \sum_{j=1}^t \varepsilon_j, \quad \tau_0 = C(1)(\mu_0 + \mu_1 / 2) + C^*(1)\mu_1, \quad \tau_1 = C(1)\mu_1 / 2.$$

Given that:

$$\mu_i = \alpha\beta_i + \alpha_{\perp}\gamma_i, \quad i=0, 1,$$

$$\beta_i = (\alpha'\alpha)^{-1}\alpha'\mu_i, \quad \gamma_i = (\alpha_{\perp}'\alpha_{\perp})^{-1}\alpha_{\perp}'\mu_i,$$

five different cases can be distinguished:

$$1) \mu_0 = \mu_1 = 0.$$

2)  $\mu_0 = \alpha\beta_0$ ,  $\mu_1 = 0$ . In this case, the constant enters the system only via the error correction term. In this case, in fact, expression (4) can be re-written as:

$$\Gamma(L) \Delta y_t = \alpha\beta'' y_{t-1}^* + \varepsilon_t, \quad \beta'' = [\beta', \beta_0], \quad y_{t-1}^* = [y_{t-1}', 1]'$$

3)  $\mu_0 = \alpha\beta_0 + \alpha_{\perp}\gamma_0$ ,  $\mu_1 = 0$ . In this model, the parameter vector on the constant is not constrained to lie in the column space of  $\alpha$ . In this case  $\mu_0$  generates a linear trend for  $y_t$ , whereas  $z_t$  has no linear trend.

4)  $\mu_0 = \alpha\beta_0 + \alpha_{\perp}\gamma_0$ ,  $\mu_1 = \alpha\gamma_1$ . In this case  $y_t$  has a linear trend, and so does  $z_t$ . The coefficient vector  $\mu_1$  lies in the column space of  $\alpha$ , and therefore the ECM representation becomes:

$$\Gamma(L) \Delta y_t = \mu_0 + \alpha\beta'' y_{t-1}^* + \varepsilon_t, \quad \beta'' = [\beta', \beta_1], \quad y_{t-1}^* = [y_{t-1}', t]'$$

5)  $\mu_0 = \alpha\beta_0 + \alpha_{\perp}\gamma_0$ ,  $\mu_1 = \alpha\beta_1 + \alpha_{\perp}\gamma_1$ . With this specification,  $y_t$  has a quadratic trend, whereas  $z_t$  has a linear trend.

To summarise, with the use of different specifications of the deterministic part of the model, it is possible to model the particular deterministic behaviour of the series under study. It is necessary to keep in mind the presence of unit roots in the

autoregressive representation. This circumstance, as in the univariate case, induces  $\mu_0$  to generate a linear trend term and  $\mu_1$  to generate a quadratic trend term. Moreover, the reduced rank nature of the matrices  $A(1)$  and  $C(1)$  causes the leading term of the deterministic trend to have different implications, depending on whether or not the associated coefficient belongs to the space spanned by the columns of  $\alpha$ . This fact is shown to have important consequences on the inferential procedures for testing for the cointegrating rank,  $r$ .

### [5.3] Estimation Issues

When interest lies in the analysis of the long run properties of potentially cointegrated vector processes, it is first necessary to assess the number of cointegrating relationships present in the system; having done this, the weights of these relationships need to be estimated.

For expositional purposes, let us assume that we have a  $(n \times 1)$  vector  $y_t$  of  $I(1)$  series, and that the cointegration rank is known and for simplicity equal to one; inference therefore focusses on the estimation of the coefficients of the  $(n \times 1)$  cointegrating vector  $\beta$ . In this respect, Engle and Granger (1987) suggest the use of a static OLS regression involving all the  $I(1)$  variables supposedly cointegrated, i.e. the  $(n \times 1)$  vector  $y_t = [y_{1t}, y_{2t}]'$ :

$$y_t = \beta' y_{2t} + z_t. \quad (8)$$

The well known result by Stock (1987) ensures that the OLS estimates are "super-consistent", in that they converge to the true parameter values at a rate  $T^{-1}$ , instead

of the rate  $T^{-1/2}$  as it happens in regressions involving stationary variables. In order to see this, assume that  $z_t$  is stationary, and that  $y_{2t}$  is generated according to:

$$\Delta y_{2t} = h_t, \quad (9)$$

where  $h_t$  is a stationary vector process. Define  $e_t = [z_t, h_t]'$ , and:

$$2\pi f_{ee}(0) = \lim_{T \rightarrow \infty} T^{-1} E \left[ \left( \sum_{t=1}^T e_t \right) \left( \sum_{s=1}^T e_s \right)' \right] = \Lambda = \Omega_0 + \Omega_1 + \Omega_1', \quad (10)$$

$$\Omega_0 = E(e_0 e_0'), \quad \Omega_1 = \lim_{T \rightarrow \infty} T^{-1} E \left[ \sum_{t=1}^T \sum_{s=2}^T e_t e_s' \right]$$

where  $f_{ee}(0)$  is the spectral density function of  $e_t$  calculated at frequency  $\omega = 0$ . Notice that the error term in the cointegrating regression and in the *DGP* for  $y_{2t}$  are both autocorrelated and cross correlated. By exploiting the usual invariance principle and the continuous mapping theorem (see Phillips and Durlauf, 1986), it is possible to show that:

$$T(\hat{\beta} - \beta) = \left( T^{-2} \sum_{t=1}^T y_{2t} y_{2t}' \right)^{-1} \left( T^{-1} \sum_{t=1}^T y_{2t} z_t \right) \quad (11)$$

$$\Rightarrow \left( \int_0^1 B_2(u) B_2(u)' du \right)^{-1} \left( \int_0^1 B_2(u) dB_1 + \Lambda_{12} \right),$$

where  $B(u)$  indicates a  $n$ -dimensional vector Brownian motion process with covariance matrix equal to  $\Lambda_{22}$ . This result ensures convergence at a rate  $T$  of the *OLS* estimator.

The well-known problem of this estimation procedure is that simulation studies (see Banerjee *et al.*, 1986) have shown that the finite distributions of the *OLS* estimates have substantial bias, persisting even in sample sizes of 100 or over.

For this reason, Phillips and Hansen (1990) propose to subject the *OLS* estimates to non-parametric corrections in order to mitigate the extent of the finite sample

bias. Their estimator is termed *Fully Modified (FM) OLS* estimation and is obtained as follows:

$$\hat{\beta}_{FM} = \left( \sum_{t=1}^T \mathbf{y}_{2t} \mathbf{y}_{2t}' \right)^{-1} \left( T^{-1} \sum_{t=1}^T \mathbf{y}_{2t} y_{1t}^* - T \hat{\Gamma}_{FM} \right), \quad (12)$$

$$y_{1t}^* = y_{1t} - \hat{\lambda}_{12} \hat{\Lambda}_{22}^{-1} \Delta \mathbf{y}_{2t}, \quad \hat{\Gamma}_{FM} = \hat{\Psi} \begin{bmatrix} 1 \\ -\hat{\Lambda}_{22}^{-1} \hat{\lambda}_{21} \end{bmatrix}, \quad \Psi = \sum_{j=0}^{\infty} E[\mathbf{h}_0 \mathbf{e}_j'],$$

where the  $\hat{\cdot}$  symbol over a variable denotes a consistent estimator of the corresponding theoretical magnitude. These estimates are obtained non-parametrically via usual kernel methods (see Newey and West, 1987, Andrews, 1991, Andrews and Monahan, 1992) starting from the residuals of the *OLS* estimate.

The effectiveness of these non parametric corrections is explained from two different viewpoints. First of all, using  $y_{1t}^*$  instead of  $y_{1t}$  is intended to reduce the effect of the long-run simultaneity, and the use of the correction  $-T \hat{\Gamma}_{FM}$  serves to reduce the effect of the "second order" bias, i.e. the bias induced by the autocorrelation properties of the error term  $\mathbf{e}_t$ . Under a more heuristic point of view, the non-parametric corrections allow use of the information contained in the *DGP* for  $\mathbf{y}_{2t}$  in order to estimate  $\beta$ . This is going to reduce the extent of the bias.

The asymptotic distribution of the *FM-OLS* estimator is obtained as:

$$T(\hat{\beta} - \beta) \Rightarrow \left( \int_0^1 \mathbf{B}_2(u) \mathbf{B}_2(u)' du \right)^{-1} \left( \int_0^1 \mathbf{B}_2(u) dB_{1,2} \right), \quad (13)$$

$$B_{1,2}(u) = B_1(u) - \lambda_{12} \Lambda_{22}^{-1} \mathbf{B}_2(u).$$

The estimation methods being surveyed so far avoid facing the important issue of how to determine the cointegrating rank. The only approach that is capable of giving this problem a sensible solution is the one developed by S. Johansen (see Johansen 1988, 1991, 1995a, 1995b, Johansen and Juselius, 1990). Johansen's

approach consists of the maximum likelihood analysis of the cointegrated *VAR* system, where inference is carried out via maximisation of the log-likelihood function. Henceforth Johansen's procedure will be referred to as *MLA* (Maximum Likelihood Approach).

In order to be able to write the likelihood function is clearly necessary to specify a joint distribution for the error vector  $\epsilon_t$ . Then, the most natural choice would be to consider  $\epsilon_t$  as multivariate normal white noise:

$$p([\epsilon_1', \epsilon_2', \dots, \epsilon_T']) = (2\pi |\Sigma|)^{-T/2} \exp\left[-\frac{1}{2} \sum_{t=1}^T \epsilon_t' \Sigma^{-1} \epsilon_t\right] \quad (14)$$

Using the *ECM* parameterisation (4), it is then possible to obtain the log-likelihood function for a finite sample of observations on  $y_t$ ,  $t = 1, \dots, T$ , conditional on the first  $k$  observations ( $y_{1-k}, \dots, y_0$ ):

$$\log L(\alpha, \beta, \Sigma, \Gamma_1, \dots, \Gamma_{k-1}, \mu) = c - (T/2) \log(|\Sigma|) - \frac{1}{2} \sum_{t=1}^T \epsilon_t' \Sigma^{-1} \epsilon_t,$$

$$\epsilon_t = \Gamma(L) \Delta y_t - \mu_0 - \alpha \beta' y_{t-1}$$

The log-likelihood maximisation strategy suggested by Johansen is based upon consecutive concentrations of the objective function. At a first step, the log-likelihood is concentrated with respect to the parameters  $\mu$ ,  $\Gamma_1$ ,  $\Gamma_2$ , ...,  $\Gamma_{k-1}$ , yielding:

$$\log L_1(\alpha, \beta, \Sigma) = c_1 - (T/2) \log(|\Sigma|) - \frac{1}{2} \sum_{t=1}^T (R_{0t} - \alpha \beta' R_{1t})(R_{0t} - \alpha \beta' R_{1t})', \quad (15)$$

where  $R_{0t}$  and  $R_{1t}$  are, respectively, the residuals of the OLS regressions of  $\Delta y_t$  and  $y_{t-1}$  on a constant and the first  $k-1$  lags of  $\Delta y_t$ . From the operative point of view, remember that this first step of the procedure is defined according to the deterministic components being allowed in the model. The case discussed here

corresponds to the most widely used model (3), when there is an unrestricted intercept term in the *ECM* representation. In other cases, one would have to define in different ways these preliminary regressions. For instance, dealing with model (2), where  $\mu_0 = \alpha\beta_0$ ,  $\Delta y_t$  and  $y_{t-1}^* = [y_{t-1}', 1]'$  are regressed on the first  $k-1$  lags of  $\Delta y_t$ .

At the second step, the log-likelihood is concentrated with respect to  $\alpha$ :

$$\log L_2(\beta, \Sigma) = c_2 - (T/2) \log(|\Sigma|) - \frac{1}{2} \sum_{t=1}^T (R_{0t} - \hat{\alpha}\beta' R_{1t})(R_{0t} - \hat{\alpha}\beta' R_{1t})',$$

$$\hat{\alpha} = S_{01}\beta(\beta'S_{11}\beta)^{-1}S_{01}, S_{ij} = T^{-1} \sum_{t=1}^T R_{it}R_{jt}', i, j = 0, 1. \quad (16)$$

Next, the function is concentrated with respect to  $\Sigma$ :

$$\log L_3(\beta) = c_3 - (T/2) \log |S_{00} - S_{01}\beta(\beta'S_{11}\beta)^{-1}\beta'S_{10}|. \quad (17)$$

Given the usual partitioned matrices results, maximising the above function with respect to  $\beta$  amounts to minimising the ratio:

$$|(\beta'S_{11}\beta) - \beta'S_{10}S_{00}^{-1}S_{01}\beta| / |(\beta'S_{11}\beta)|. \quad (18)$$

This context is very similar to the *LIML* estimation approach (see for instance Davidson and Mc Kinnon, 1994, pp. 644-651). It is therefore possible to work with the normalisation  $\beta'S_{11}\beta = I_r$  and show that the  $(n \times r)$  matrix  $\beta$  which minimises (18) is given by taking the  $r$  generalised eigenvalues of  $S_{10}S_{00}^{-1}S_{01}$  with respect to  $S_{11}$ , corresponding to the  $r$  largest eigenvalues. The maximum of the log-likelihood function is therefore:

$$\begin{aligned}\log L^*(r) &= c - T/2 \log |\hat{\beta}_r' S_{11} \hat{\beta}_r - \hat{\beta}_r' S_{10} S_{00}^{-1} S_{01}| = \\ &= c - T/2 \log |I_r - \hat{\Lambda}_r| = c - T/2 \sum_{i=1}^r \log(1 - \hat{\lambda}_i),\end{aligned}\quad (19)$$

where  $\hat{\Lambda}_r$  is a  $(r \times r)$  diagonal matrix with the  $r$  largest generalised eigenvalues on its main diagonal.

It is possible to provide a different interpretation to the *MLA* estimator. In fact, as stressed in Johansen (1988), the estimates of  $\beta$  and  $\alpha$  are related to the canonical variates between  $R_{0t}$  and  $R_{1t}$  (see Anderson, 1984): the *ML* estimate of  $\beta$  corresponds to the  $r$  linear combinations of  $y_{t-1}$  having the largest squared partial correlations with  $\Delta y_t$ , after having corrected for the effects of the variables appearing as regressors in the preliminary regressions. This interpretation of the estimates is based on the nature of reduced rank regression of the *ECM* representation.

On the basis of these results, it is possible to construct a likelihood ratio test in order to test  $H_0$ : cointegration rank =  $r$  against the alternative  $H_1$ : cointegration rank =  $n$ :

$$LR(r/n) = -T \sum_{i=r+1}^n \log(1 - \hat{\lambda}_i), \quad (20)$$

and this test is known as the *trace* test. In the same way, it is possible to obtain the likelihood ratio test in order to test  $H_0$ : cointegration rank =  $r$  against the alternative hypothesis  $H_1$ : cointegration rank =  $r+1$ :

$$LR(r/r+1) = -T \log(1 - \hat{\lambda}_{r+1}), \quad (21)$$

known as  $\lambda$ -max test.

The finite sample distributions of these statistics are completely unknown, but the asymptotic properties have been deeply analysed (see for instance chapter 11 in Johansen 1995a). For ease of exposition, let us concentrate only on the *trace* test. It is possible to show that the following result holds:

$$-T \sum_{i=r+1}^n \log(1 - \hat{\lambda}_i) \Rightarrow \text{trace} \left\{ \int_0^1 (d\mathbf{W}) \mathbf{F}' \left[ \int_0^1 \mathbf{F} \mathbf{F}' d\mathbf{W} \right]^{-1} \int_0^1 \mathbf{F} (d\mathbf{W})' \right\}, \quad (22)$$

where  $\mathbf{W}$  denotes a standard Brownian motion process in  $p-r$  dimensions, and  $\mathbf{F}$  is a function of  $\mathbf{W}$  defined in different ways depending on the particular deterministic part of the model. Recalling the five different models described above:

- 1) When  $\mu_0 = \mu_1 = 0$ ,  $\mathbf{F}(u)$  coincides with  $\mathbf{W}(u)$ .
- 2) When  $\mu_0 = \alpha\beta_0$ , and  $\mu_1 = 0$ ,  $\mathbf{F}(u)$  has  $p-r+1$  dimensions and we have:

$$F_i(u) = W_i(u), i = 1, 2, \dots, p-r,$$

$$F_i(u) = u, i = p-r+1.$$

- 3) When  $\mu_0 = \alpha\beta_0 + \alpha_1\gamma_0$ , and  $\mu_1 = 0$ , we have:

$$F_i(u) = W_i(u) - \int W_i(u) du, i=1, 2, \dots, p-r-1,$$

$$F_i(u) = u-1/2, i=p-r.$$

- 4) When  $\mu_0 = \alpha\beta_0 + \alpha_1\gamma_0$  and  $\mu_1 = \alpha\beta_1$ , the  $\mathbf{F}(u)$  process in (22) is  $p-r+1$ -dimensional, and is defined as:

$$F_i(u) = W_i(u) - \int W_i(u) du, i=1, 2, \dots, p-r,$$

$$F_i(u) = u-1/2, i=p-r+1.$$

- 5) Finally, when both  $\mu_0$  and  $\mu_1$  are unconstrained, the  $\mathbf{F}(u)$  process has  $p-r$  dimensions and it is defined as:

$$F_i(u) = W_i(u) - a_i - b_i u, i=1, 2, \dots, p-r-1,$$

$$F_i(u) = u^2 - a - b u, i=p-r,$$

where the coefficients  $a_i$ ,  $b_i$ ,  $a$  and  $b$  are obtained by regressing respectively  $W_i(u)$  and  $u^2$  on an intercept and a linear trend.

If the deterministic part of the model were different from any of the five cases described above, the asymptotic distribution results could be radically different. Everything depends on which term asymptotically dominates the deterministic behaviour of the process. For instance, the presence of an intercept-shifting dummy



variable would modify the asymptotic distributions of the cointegrating rank statistics in case 3, i.e. when the leading deterministic term is the constant term, but it would not change anything in case 5, where the leading deterministic term is a linear trend.

In synthesis, dependence of the asymptotic distribution (22) on the deterministic part of the model renders inference somehow problematic. Exactly as happens in univariate unit root testing, we need to determine correctly the deterministic features of the model, in order to conduct correct inference on the stochastic features of the series under study. Hence the inferential results are somehow conditional on the choice of the deterministic component being valid.

Ironically, the restrictions associated with each of the different deterministic components described above could be tested by means of a standard asymptotically  $\chi^2$ -distributed *LR* test, *given* the cointegrating rank, as we will see when dealing with the distributional properties of the estimates. The implicit circularity of the procedure is evident.

In order to cope with the problem, Johansen (1992) follows Berger and Sinclair (1984) and Pantula (1989) and specifies an approach based on testing a nested sequence of hypotheses. The main idea behind this approach is to reject an hypothesis only if the hypotheses contained in it are rejected. For instance, let us suppose that it is not clear whether to adopt model (2) or model (3) as the best description of the deterministic feature of the data. Defining  $H_i(r)$  as the rank  $r$  hypothesis in model  $i$  ( $=2$  or  $3$ ), and  $c_i(r)$  the  $\alpha\%$  quantile of the asymptotic distribution of the corresponding *trace* test statistic  $Q_i(r)$ , Johansen proposes to reject  $H_i(r)$  if the collection of test results for all the contained hypotheses belong to the set:

$$\{Q_h(k) > c_h(k), \forall h, k \text{ such that } H_h(k) \subseteq H_i(r)\},$$

and to accept  $H_i(r)$  if the collection of test results for all the contained hypotheses belong to the set:

$$\{Q_h(k) > c_h(k), \forall h, k \text{ such that } H_h(k) \subset H_i(r), \text{ and } Q_i(r) < c_i(r)\}.$$

This testing procedure consists in testing a sequence of hypotheses where the hypotheses further on in the sequence contain all the preceding ones. Johansen (1992) shows that this procedure is consistent and it has asymptotic size equal to  $\alpha$ . Of course very little is known about the finite sample properties of this testing procedure: "The inference conducted here is asymptotic and simulations show that one can easily find situations in practice where the number of observations is not sufficient to apply asymptotic results" (Johansen, 1995a, chapter 11).

#### [5.4] Interpretation of the Cointegrating Coefficients

As stressed in the previous section, the cointegrated *ECM* model is affected by a lack-of-identification problem. In Johansen's maximum likelihood approach this lack-of-identification problem is solved by adopting the normalisation  $\beta'S_{00}\beta = I_r$ . Of course, this normalisation does not necessarily have any economically meaningful interpretation.

Johansen and Juselius (1994) and Johansen (1995b) give a solution to the problem of the interpretation of the cointegration relationships by casting it into a classical identification problem. The cointegration relationships  $\beta'y_t = z_t$  can be interpreted as a system of  $r$  linear equations. In order to achieve identification, it is possible to impose a set of constraints on each equation. A set of  $r$  normalisation constraints is needed in order to impose a unit coefficient on one of the variables in each

equation. Leaving these trivial constraints aside, Johansen considers linear homogeneous constraints of the kind :

$$\mathbf{R}_i' \boldsymbol{\beta}_i = 0, \quad i=1, 2, \dots, r, \quad (23)$$

where  $\boldsymbol{\beta}_i$  indicates the  $i$ -th column of  $\boldsymbol{\beta}$  and  $\mathbf{R}_i$  is a  $(r \times q_i)$  full column rank matrix. The same constraints can be expressed in explicit form as follows:

$$\boldsymbol{\beta}_i = \mathbf{H}_i \boldsymbol{\phi}_i, \quad i=1, 2, \dots, r, \quad \mathbf{R}_i' \mathbf{H}_i = 0. \quad (24)$$

Following Sargan (1988), identification of the  $i$ -th equation is achieved when the following rank condition is fulfilled:

$$\rho(\mathbf{R}_i' \boldsymbol{\beta}) = r-1,$$

meaning that the "structural"  $i$ -th equation, i.e. the one obeying the constraints (23) cannot be generated as a linear combination of the other columns of  $\boldsymbol{\beta}$ . The rank condition is satisfied only when the order condition  $q_i \geq r-1$  is fulfilled.

Nevertheless, it is problematic to check the rank condition because it impinges upon the values of unknown parameters. For this reason, Johansen (1995b) puts forward another formulation of the rank condition which is entirely based upon the structure of all the constraints being imposed upon  $\boldsymbol{\beta}$ . The constraints imposed on the system are such to identify the  $i$ -th equation if and only if:

$$\rho\left\{\mathbf{R}_i' \left[ \mathbf{H}_{j_1} | \mathbf{H}_{j_2} | \dots | \mathbf{H}_{j_k} \right] \right\} \geq k,$$

for every set  $j_k, 1 \leq j_1 < j_2 < \dots < j_k \leq r$  with  $k=1, 2, \dots, r-1$ . If the  $i$ -th equation is identified and the rank condition is satisfied as  $q_i = r-1$ , then the equation is *exactly* identified, and no constraint is actually being imposed on it. If, instead, the  $i$ -th equation is identified and  $q_i > r$ , the equation is *overidentified* and  $q_i - r + 1$  constraints are actually being imposed on it. In order to fully understand this concept, let us

consider Phillips's triangular representation as a particular example of exact identification. The particular structure being imposed on the cointegrating relationships complies with Johansen's identification conditions, and it does not entail any constraint being imposed on the parameter space; in fact, leaving aside the normalisation constraint forcing the  $i^{\text{th}}$  variable to appear with a unit coefficient in the  $i^{\text{th}}$  equation ( $i=1, 2, \dots, r$ ), on each equation we have  $r-1$  exclusion restrictions: all the equations are exactly identified and no restriction is being imposed on the parameter space. From a different viewpoint, this finding is corroborated by the fact that the triangular representation can be obtained by simple algebraic transformation of the unrestricted estimation which is obtained subject to the normalisation  $\beta'S_{11}\beta = I_r$ .

When over-identifying constraints are being imposed, their legitimacy can be tested by means of Wald or likelihood ratio test statistics. Using the *LR* testing principle requires estimation of the model subject to the restrictions. This can be achieved by means of a *switching* algorithm which works as follows: starting from an arbitrary initialisation, one cyclically solves the reduced rank regression algorithm for each one of the columns of  $\beta$ , imposing all the constraints and considering all the other columns of  $\beta$  as given. This is shown to converge to the maximum likelihood estimation under the hypothesis that the over-identifying constraints hold. A *LR* test can be easily constructed, and Johansen (1995b) shows that the resulting statistic is asymptotically  $\chi^2$  distributed with as many degrees of freedom as the number of over-identifying restrictions being imposed on the parameter space.

### [5.5] Asymptotic Distributions of the Parameter Estimates

After having adopted a normalisation to identify the cointegration relationships parameters (for instance the triangular representation normalisation), the asymptotic distribution of the estimates can be obtained by Taylor expansion of the log-likelihood function. For the sake of brevity, I only deal with the case in which the deterministic part is equal to  $\mu_0$ . The asymptotic distributions applying in all the other cases can be easily obtained by appropriately modifying the Brownian motion processes involved.

Working with the normalisation  $C'\hat{\beta} = C'\beta = I_r$ ,  $C' = [I_r, 0]$ , which corresponds to the triangular representation, theorem 12.3 of Johansen (1995a) proves the following result :

$$\begin{aligned} T(\hat{\beta} - \beta) &\Rightarrow (I_n - \beta C') \bar{\gamma} \left[ \int_0^1 G_{12} G_{12}' du \right]^{-1} \int_0^1 G_{12} (dV_\alpha)', \\ G_{12} &= G_1 - \int_0^1 G_1 G_2' du \left[ \int_0^1 G_2 G_2' du \right]^{-1} G_2, \\ G_1 &= \bar{\gamma}' C(1) \left[ W(u) - \int_0^1 W(u) du \right], \quad G_2 = u - 1/2, \\ V_\alpha &= (\alpha' \Sigma^{-1} \alpha)^{-1} \Sigma^{-1} W(u), \quad \bar{\gamma} = \gamma(\gamma' \gamma)^{-1}, \end{aligned} \quad (25)$$

$W(u)$  in this context indicates a vector Brownian motion with covariance matrix equal to  $\Sigma$ , and  $\gamma$  is a  $(n \times (s-1))$  matrix orthogonal to  $\beta$  and to the coefficient vector of the leading deterministic term in the Beveridge-Nelson representation.

This result tells us two things:

- 1) As in the static regression, the normalised coefficients in  $\beta$  are *super-consistent*, since they converge at a rate  $T^{-1}$  to their true values;
- 2) the asymptotic distribution of  $Tvec(\hat{\beta} - \beta)$  is *mixed-Normal*, with mixing covariance matrix given by:

$$[(\alpha' \Sigma^{-1} \alpha)^{-1}] \otimes \left\{ (\mathbf{I}_n - \beta' \mathbf{C}) \bar{\gamma} \left[ \int_0^1 \mathbf{G}_{12} \mathbf{G}_{12}' du \right]^{-1} \bar{\gamma}' (\mathbf{I}_n - \beta' \mathbf{C}) \right\},$$

which can be consistently estimated as:

$$T[(\hat{\alpha}' \Sigma^{-1} \hat{\alpha})^{-1}] \otimes \left\{ (\mathbf{I}_n - \hat{\beta}' \mathbf{C}) \mathbf{S}_{11}^{-1} (\mathbf{I}_n - \hat{\beta}' \mathbf{C}) \right\},$$

given that we will see that  $\hat{\alpha}$  is a consistent estimator for  $\alpha$ .

The circumstance that  $\hat{\beta}$  is asymptotically mixed-Normal means that, conditionally on the estimate of the mixing covariance matrix, it is possible to use the standard distribution theory which will be asymptotically valid. Nevertheless, one has to take into account that the marginal asymptotic distribution of  $\hat{\beta}$  will have fatter tails than a Normal distribution.

When over-identifying constraints such as the ones described in expression (24) are imposed on the cointegrating vectors, and a unit normalisation has been imposed such that the constraints become:

$$\beta_i = \mathbf{H}' \phi_i + \mathbf{h}_i, \quad sp(\mathbf{h}_i, \mathbf{H}') = sp(\mathbf{H}_i), \quad i = 1, 2, \dots, r, \quad \mathbf{R}_i' \mathbf{H}_i = \mathbf{0},$$

the asymptotic distribution of  $vec(\hat{\beta})$  is still mixed-Normal, but the mixing covariance matrix is different from the one described above, and it can be consistently estimated as:

$$T\{\mathbf{H}'\} \left\{ (\hat{\alpha}, \hat{\Sigma}^{-1} \hat{\alpha}) \mathbf{H}'' \mathbf{S}_{11} \mathbf{H}' \right\}^{-1} \{\mathbf{H}''\}.$$

In the above expression the notation  $\{\mathbf{A}_i\}$   $i=1, 2, \dots, r$ , indicates a block diagonal matrix with  $i^{\text{th}}$  block equal to  $\mathbf{A}_i$ , and  $\{\mathbf{A}_{ij}\}$ ,  $i, j=1, 2, \dots, r$ , means a partitioned matrix with blocks  $\mathbf{A}_{ij}$ .

The super-consistency property of the cointegrating vector coefficients allows one to obtain the asymptotic distributions of the estimates of the loading factors and of the parameters connected to the short run dynamics. In fact, given super-consistency of  $\hat{\beta}$ ,  $\beta$  can be asymptotically considered as known. Writing the *ECM* representation (4) as:

$$\begin{aligned}\Delta \tilde{y}_t &= \Xi' \tilde{h}_t + \varepsilon_t, \\ \Xi &= [\alpha | \Gamma_1 | \dots | \Gamma_{k-1}], \quad h_{it} = [y_{t-1}' \beta | \Delta y_{t-1}' | \dots | \Delta y_{t-k+1}']',\end{aligned}\tag{26}$$

where the  $\sim$  symbol over a variable means the residuals of a regression of that variable on the unrestricted deterministic part of the *ECM* model, it is clear that, were  $\beta$  known, all the variables appearing in (26) would be stationary. For this reason, standard asymptotic results apply for the parameters in  $\Xi$ :

$$\begin{aligned}T^{1/2} \text{vec}[\hat{\Xi} - \Xi] &\xrightarrow{w} N(0, \Sigma \otimes \Omega^{-1}), \\ \Omega &= \text{var}(h_t).\end{aligned}$$

This allows the use of standard asymptotic results also on the parameters of the *VAR* representation, since these are linear functions of the elements of  $\Xi$ , considering  $\beta$  as given. This fact is particularly useful in order to obtain the asymptotic distributions of the impulse response functions and of the forecast error variance decompositions in a cointegrated *VAR* model, which are continuous non-linear functions of the autoregressive parameters.

### [5.6] Finite Sample Properties

In the previous section I have reviewed the asymptotic distributions of the maximum likelihood estimate of the long-run parameters in a cointegrated *VAR*

model. These results are the basis of all the inferential procedures being widely used in the applied literature. Blind reliance on asymptotic results can be dangerous, given that in the typical macroeconomic application it is very common to work with very short sample periods. Therefore, it is extremely useful to investigate small sample properties of the *ML* estimates and, more generally, the properties of the test statistics being used to guide key decisions concerning the specification of the model.

From the theoretical point of view, a recent paper by P.C. B. Phillips (1994) has investigated the exact finite sample distributions of the normalised reduced rank estimates of the cointegrating parameters. The analytical results obtained by Phillips echo the analogous results concerning the exact finite sample distribution of *LIML* estimates in a simultaneous equation model (Phillips, 1983), and this is not surprising given the already mentioned analogy between the *ML* estimate of  $\beta$  and the *LIML* estimator.

Analysing a simple *ECM* model as in (4) but without deterministics, and working with the normalised estimates  $\hat{\beta} = [I, \hat{\phi}]'$ , Phillips discovers that the leading term of the finite sample distribution of  $\hat{\phi}$  is proportional to  $|I_r + \hat{\phi}'\hat{\phi}|^{-n/2}$ , i.e. to the kernel of a matrix-Cauchy distribution. The exact shape of the distribution is in general very complicated, but its tail behaviour is generated by the matrix-Cauchy term. This feature means that the finite sample distribution does not have finite moments of integer order. From a different viewpoint, the free parameters in the normalised estimates are obtained as the ratio of two blocks of the *ML* estimates; as stressed in Sargan (1988) and Phillips (1983), this is enough to prevent the finite sample distribution of the resulting coefficients from having finite moments of integer order. As Phillips (1994) emphasises, the Cauchy-like tail behaviour is therefore not a consequence of the circumstance that the asymptotic distribution of  $\hat{\phi}$  is mixing-normal, i.e. that in the limit the sample information is random: in fact,



Phillips' (1991b) estimator based on the triangular representation does not have Cauchy-like tails. However, it is necessary to keep in mind that this estimator does not allow to conduct inference on the cointegrating rank.

Some Monte Carlo simulation studies shed further light on the finite sample properties of cointegrating coefficient estimators. Cappuccio and Lubian (1995) conduct a very interesting experiment simulating a six-variate cointegrated *DGP* with rank equal to two and subject to over-identifying restrictions on the cointegrating parameters. They generate 10,000 samples of data according to:

$$y_{1t} = \Phi' y_{2t} + u_{1t}, \Delta y_{2t} = u_{2t}, u_t = A u_{t-1} + \varepsilon_t, \varepsilon_t \sim N.i.d(0, \Sigma), \quad (27)$$

where  $y_{1t}$  is bivariate and  $y_{2t}$  is 4-variate, choosing a certain value for  $\Phi$  and a range of different values for  $A$  and  $\Sigma$ .

Then they estimate a *ECM* model subject to the over-identifying restriction for each data set and for different sample sizes. At each step also the tests of the over-identifying constraints are computed. Their results can be summarised as follows.

a) The reduced rank regression *ML* estimates do not show significant finite sample bias, but they have huge numerical standard errors, reflecting the Cauchy-like tails of the finite sample distributions. The occurrence of outliers in the estimates becomes negligible only for sample sizes of 200. This means that the any applied macroeconomic researcher should be extremely careful in relying on the asymptotic distributions of the parameter estimates.

b) Even more alarmingly, the finite sample distributions of the *LR* test used to check the validity of the over-identifying restrictions are very different from their asymptotic  $\chi^2$  counterparts. For a sample size of 50, and high values of the time dependency of the  $u_t$  process, the actual size of the testing procedure applied with a nominal size of 5% is almost 80%, clearly leading to extreme over-rejection of

the null hypothesis. Only with sample sizes equal to 300, does the actual test size tend to become close to the nominal one.

These simulation results clearly signal that the finite sample properties of the *MLA* inferential procedures can be substantially different from their asymptotic counterparts. In my view, this is already enough to justify the quest for a different inferential strategy, based on exact finite sample results, in order to avoid reliance on incorrect distributional theory. This is the main rationale behind the use of Bayesian inferential techniques in the analysis of cointegrated system carried out in the next chapter.

## **Chapter 6: Bayesian Inference in Cointegrated Systems**

### **[6.0] An Overview of the Chapter.**

In this chapter I develop a Bayesian procedure to conduct inference on the cointegrating rank of a system of  $I(1)$  variables, and to verify the validity of over-identifying restrictions on the cointegration parameters. The model is specified in terms of a parameterisation which seems to suit well this inferential problem. Exact finite sample distributions for the parameters and the statistics of interest are obtained by means of Monte Carlo integration of the corresponding conditional posterior distributions. A simulation analysis, an application on Danish and Finnish money demand data, and an application on the UK exchange rate data are presented. The chapter is organized as follows. Recalling the content of chapter [5], Section [6.1] summarises the main motivations behind a Bayesian approach to the analysis of cointegrated systems. Section [6.2] is devoted to presenting the model. Section [6.3] describes the prior distribution and Section [6.4] copes with the resulting joint distribution. Section [6.5] considers inference on the cointegration rank, and Section [6.6] deals with testing the validity of the over-identifying restrictions imposed on the cointegrating vectors. Section [6.7] contains the results of a set of applications. The applications are the Danish and Finnish money demand examples studied by Johansen and Juselius (1990), and the PPP /UIP UK data of Johansen and Juselius (1992). Section [6.8] concludes.

### [6.1] Motivations

As I have detailed in the previous chapter, Johansen's *ML* approach to the analysis of cointegrated systems presents some technical problems which can be summarised in the following few points:

1) The cointegrating rank test statistics have non standard distributions, and only asymptotic results are available for any kind of inference in the model. These distributions have to be tabulated, by simulating the vector Brownian motion process a sufficiently high number of times. Once the rank has been determined on the basis of the relevant statistics, since the parameters of the cointegrating vectors are not identified, the researcher has to provide the model with some (linear) constraints intended to achieve identification. Insofar as these constraints generate over-identification of the cointegrating parameters, their validity can be checked by means of a likelihood ratio test statistic (see Johansen, 1995b), whose distribution is standard (asymptotically  $\chi^2$ ). All the inferential questions in the model, conditioned upon a given cointegrating rank, can be answered on the basis of available standard distributional results. The identified cointegrating vectors themselves can be given a distribution (Johansen 1991, 1995a, 1995b): the asymptotic distribution is normal, with a variance-covariance matrix which is itself a random variable. Therefore the identified cointegrating vectors have an asymptotic distribution which is in the form of a mixture of normal variables. There is also the necessity of resorting to asymptotic results when the interest of the researcher focuses on non-linear functions of the parameters of the model, e.g. the impulse response coefficients (see Lutkepohl 1991) or the shock persistence profiles (see Pesaran and Shin, 1995). As for the finite sample distributions of the parameters, little is known, beyond Phillips' (1994) finding that identified cointegrating vectors have finite sample Cauchy-like tailed distributions. Cappuccio

and Lubian (1995) provide disturbing Monte Carlo evidence on the poorness of finite sample properties of the *LR* tests of the over-identifying restrictions.

2) An awkward role is played by deterministic components. As pointed out in Section [5 6], different deterministic components generate different asymptotic distributions for the cointegration rank test statistics. Exactly as happens in conventional unit root testing, there is the usual necessity to decide what kind of deterministic behaviour to allow for the model being used to conduct inference on the stochastic features of the series under study. Then the inference results are somehow conditional on the deterministic component being allowed for. In order to cope with the problem, Johansen (1992) has specified a procedure whose finite sample properties are unknown.

3) Inefficient parameterisation. The test-bed parameterisation for the presence of cointegrating long-run relationships is the vector autoregressive (*VAR*) framework. In Sims' (1980) own words, this is a "profligate" parameterisation, which precludes the possibility of analysing models with more than 4 or 5 series. Imposing constraints on the parameter space is not easily feasible in Johansen's approach. Restricting the parameter space would preclude using simple OLS partial regression to concentrate the likelihood function, as is done in Johansen's setting. These issues justify an alternative approach to the problem. The Bayesian approach seems to be the obvious candidate, if one considers all the points mentioned above. As for point one, Bayesian inferential techniques are very much more straightforward in their applied interpretation, and they easily lend themselves to become a convenient support to decision making in modelling. Moreover they are based on exact finite sample distributions of the relevant parameters and statistics. This might prove of crucial importance, as an interesting study by Bauwens and Lubrano (1994) has recently shown: if one were to accept the asymptotic standard error associated with the identified cointegration coefficient vectors as a true

measure of the uncertainty associated with the results being obtained, one would actually underestimate that uncertainty. The exact finite sample posterior distribution is much more dispersed. Using a Bayesian framework of analysis, one could compute the moments and the exact finite sample posterior distribution of any desired function of the parameters.

As for the second point, instead of conditioning upon the deterministic component, as is implicitly done in the *MLA*, using Bayesian techniques one could think of marginalising the joint posterior distribution with respect to the parameters in the deterministic part. This would really render the analysis more coherent.

As for the problem of parameterisation inefficiency, once again the use of Bayesian techniques might render it possible to resort to some useful and sensible ways to restrict the parameter space. For example, one might consider the *BVAR*-type of approach (see Doan, Litterman and Sims, 1984), where one trades off unbiasedness with efficiency, or alternative specifications of distributed lags.

Of course many computational difficulties are expected to arise, and indeed do arise. Nevertheless, many different numerical simulation techniques are nowadays available, as shown in Chapter 3. In the present study I use Monte Carlo integration techniques via Gibbs sampling in order to conduct posterior inference in a cointegrated VAR setting. The inference being conducted refers to the number of cointegrating vectors and to their structural interpretation, and this constitutes the main novelty of this approach with respect to previous Bayesian studies: for example Kleibergen and van Dijk (1994) work with a given cointegrating rank.

## **[6.2] The Model**

Consider the following *VAR* model for  $n$ -dimensional  $I(1)$  vector series  $y_t$  :

$$A(L)y_t = \delta' D_t + \varepsilon_t, \quad A(L) = I_n - \sum_{j=1}^k A_j L^j,$$

or

$$\Gamma(L)\Delta y_t = \alpha\beta' y_{t-1} + \delta' D_t + \varepsilon_t, \quad \Gamma(L) = I_n - \sum_{j=1}^{k-1} \Gamma_j L^j \quad (1)$$

$$\rho(\alpha) = \rho(\beta) = r \leq n,$$

$$\varepsilon_t \sim NWN(0, \Sigma).$$

The vector  $D_t$  contains the deterministic components of the model. The parameters in  $\delta'$  are convolutions of the autoregressive parameters and the expectation value of  $y_t$ .

Model (1) is exactly the one specified by Johansen and his co-authors in his papers. The model is subject to the usual non-identification issue of the cointegrating parameters. As seen in Section [5.4], some restrictions are needed in order to achieve identification. I decided to start from the condition usually imposed, i.e. the normalisation  $\beta = [I_r, \varphi']'$ , which corresponds to a situation of exact identification of  $\beta$ .

In order to write the likelihood function of the process, I write the model in a matrix form:

$$\Delta Y = D\delta + Y^* \Gamma + Y_{-1} \beta \alpha' + E, \quad \text{vec}(E) \sim N(0, \Sigma \otimes I_T),$$

$$\Gamma = [\Gamma_1 | \Gamma_2 | \dots | \Gamma_{k-1}], \quad \{Y^*\}_t = [y_{t-1} | y_{t-2} | \dots | y_{t-k+1}] \beta$$

The likelihood function of the model is then:

$$p(\text{data} | \alpha, \beta, \Gamma, \delta, \Sigma) \propto |\Sigma|^{-T/2} \exp\{-.5 \text{trace}(E' E \Sigma^{-1})\}. \quad (2)$$

Notice that the model is bilinear in the parameters  $\alpha, \Gamma, \delta$  on one side, and  $\beta$  on the other. Therefore, the conditional distribution of each group of the parameters, given all the others, are of known analytical forms. This will be used in the

posterior distribution analysis, after having described the prior distribution being implemented.

### [6.3] The Prior Distribution.

The main aim behind the specification of the prior distribution is to ensure enough flexibility to represent the extra sample information being available to the researcher, with the desired strength, as measured by prior precision. The prior distribution is denoted as:

$$p(\alpha, \phi, \Gamma, \delta, \Sigma) = p(\alpha) p(\beta) p(\Gamma) p(\delta) p(\Sigma).$$

Note the assumption of prior independence among the set of parameters  $\alpha$ ,  $\phi$ ,  $\Gamma$ ,  $\delta$ ,  $\Sigma$ . This assumption is by no means necessary and can be relaxed only at the cost of making computations slightly more burdensome.

The prior pdfs for  $\alpha$  and  $\phi$  read:

$$\begin{aligned} p(\text{vec}(\alpha)) &\propto \exp\{-.5[(\text{vec}(\alpha) - \mu_\alpha)' \Sigma_\alpha^{-1} (\text{vec}(\alpha) - \mu_\alpha)], \\ p(\text{vec}(\phi)) &\propto \exp\{-.5[(\text{vec}(\phi) - \mu_\phi)' \Sigma_\phi^{-1} (\text{vec}(\phi) - \mu_\phi)], \end{aligned} \tag{3}$$

As in Geweke (1993), these are shrink-to-mean prior distributions:  $\mu_\alpha$  and  $\mu_\phi$  are the prior means for  $\text{vec}(\alpha)$  and  $\text{vec}(\phi)$ , respectively. Note that in many macroeconomic applications, economic theory suggests prior beliefs on the long-run equilibrium relationships. These beliefs can easily be reflected with an adequate choice of the hyperparameter vector  $\mu_\phi$ . The intensity of these prior beliefs is



proportional to  $\Sigma_{\alpha}^{-1}$  and  $\Sigma_{\phi}^{-1}$ . For convenience, I choose the prior variance matrices to be:

$$\Sigma_{\alpha}^{-1} = \omega_{\alpha} \mathbf{I}_{nr}, \quad \Sigma_{\phi}^{-1} = \omega_{\phi} \mathbf{I}_{rs}, \quad \text{where } s = n-r. \quad (4)$$

The single hyperparameters  $\omega_{\alpha}$  and  $\omega_{\phi}$  control the strength of prior beliefs on  $\alpha$  and  $\phi$  respectively. The necessity of specifying a proper prior distribution for the parameters in  $\phi$  will be justified when describing the joint posterior pdf.

The prior distribution of  $\text{vec}(\Gamma)$  is specified as:

$$p(\text{vec}(\Gamma)) \propto \exp\{-.5 \text{vec}(\Gamma)' \Sigma_{\Gamma}^{-1} \text{vec}(\Gamma)\}, \quad (5)$$

where  $\Sigma_{\Gamma}^{-1}$  is a diagonal matrix whose elements are determined as follows:

$$[\text{var} \{\Gamma_h\}_{ij}]^{-1} = \omega_1 \omega_2^h \omega_3 \quad \text{when } i \neq j, \quad [\text{var} \{\Gamma_h\}_{ij}]^{-1} = \omega_1 \omega_2^h \quad \text{when } i = j. \quad (6)$$

$$\omega_2^h > 0, \quad \omega_3 > 0.$$

As in Doan, Litterman and Sims (1984) a shrink-to-mean prior is specified: each autoregressive parameter has a prior mean equal to zero, with precision increasing with the lag order and, *ceteris paribus*, when the coefficient is off the diagonal of the autoregressive matrices: this amounts to believe that own lags are more important in each of the *VAR* equations, and that coefficients on relatively recent lags are expected to drift from zero more than those on relatively distant lags. This seems to be a viable and sensible solution to the overparameterisation problem related to the *VAR* setting. Of course, when the overall tightness hyperparameter ( $\omega_1$ ) is zero, the prior is diffuse.

Also for the deterministic parameters in  $\delta$ , a normal shrink-to mean-prior distribution is specified:

$$p(\text{vec}(\delta)) \propto \exp\{-.5(\text{vec}(\delta)-\mu_\delta)' \Sigma_\delta^{-1} (\text{vec}(\delta)-\mu_\delta)\}, \quad (7)$$

where a prior precision  $\Sigma_\delta^{-1}$  equal to zero reflects prior ignorance.

The prior for  $\Sigma$  is:

$$p(\Sigma) \propto |\Sigma|^{-(n+1)/2}, \quad (8)$$

and it is a customary choice in the Bayesian treatment of multi-equational models, in the absence of prior information. In case prior information about  $\Sigma$  is available, it is possible to specify an inverted Wishart (Zellner, 1971, Section B.3) specification for  $p(\Sigma)$ , of which expression (8) constitutes a special case.

In synthesis, I choose independent shrink-to-mean priors for the parameters describing the conditional mean of the process  $y_t$ , i.e.  $\alpha$ ,  $\phi$ ,  $\Gamma$ ,  $\delta$ . For  $\Sigma$  I specify a customary ignorance prior.

Note that for each of these five groups of parameters, and conditionally on the other ones, the prior pdf is conditionally conjugate with respect to the corresponding conditional likelihood. In this way, the conditional posterior distributions are all tractable.

#### **[6.4] The Joint Posterior Distribution.**

Combining the likelihood function with the prior distribution, it is possible to write the joint posterior distribution:

$$p(\alpha, \phi, \Gamma, \delta, \Sigma | data) \propto |\Sigma|^{-(T+n+1)/2} \exp \left\{ -\frac{1}{2} [\text{trace} (\mathbf{E}' \mathbf{E} \Sigma^{-1} + (\text{vec}(\alpha) - \mu_\alpha)' \Sigma_\alpha^{-1} (\text{vec}(\alpha) - \mu_\alpha) + (\text{vec}(\phi) - \mu_\phi)' \Sigma_\phi^{-1} (\text{vec}(\phi) - \mu_\phi) + \text{vec}(\Gamma)' \Sigma_\Gamma^{-1} \text{vec}(\Gamma) + (\delta - \mu_\delta)' \Sigma_\delta^{-1} (\delta - \mu_\delta)] \right\}. \quad (9)$$

Notice that, given the informative prior pdf, in the joint posterior pdf all the parameters are identified. If the prior distribution for  $\phi$  were improper, i.e. if  $\omega_\phi = 0$ , a rank deficiency in  $\alpha$  would induce singularity in the variance covariance matrix in the conditional posterior pdf of  $\text{vec}(\phi)$ . This is the point made in Kleibergen and van Dijk (1994).

Due to the already mentioned presence of non linearities between parameters, the joint posterior distribution is not easily amenable to analytical integration. Therefore, in order to obtain posterior marginal distributions and posterior moments, it is necessary to resort Monte Carlo Integration. This technique has been explained in Section [3.1].

When it is not possible to provide i.i.d. draws from the joint posterior distribution, as in our case where it is of no known analytical form, then some other methods have to be adopted. Following the suggestions of Hammersley and Handscomb (1964) one could choose an "importance function" to sample from. In any case that choice is not easy, and it might yield very poor estimates. In fact, as is stressed in Koop (1994), it is necessary that the tails of the importance distribution be fatter than those of the posterior distribution, otherwise the draws from the tails of the importance function dominate the behaviour of the Monte Carlo estimate. For this reason, one should know exactly the shape of the posterior distribution, in order to choose correctly the importance function. We do not know the form of the joint posterior in our context, and therefore we adopt a Gibbs Sampling Algorithm (GSA). This algorithm has been described in Section [3.3c]. Application of the GSA requires the possibility of obtaining random draws from the conditional

posterior distributions in the present context. The parameter vector is divided into three sets, the first one  $\theta_1 = [\text{vec}(\varphi)]$ , the second one  $\theta_2 = [\text{vec}(\alpha') | \text{vec}(\Gamma) | \text{vec}(\delta)']$ , and the third one  $\theta_3 = \text{vech}(\Sigma)$ . The following three lemmas describe the conditional posterior distributions of these three subsets of parameters.

**Lemma 4.1:** The conditional posterior distribution of  $\theta_1 = \text{vec}(\varphi)$  is  $r \times s$ -variate normal, with moments:

$$E(\text{vec}(\varphi) | \alpha, \Gamma, \delta, \Sigma \text{ data}) = [Q_d + Q_p]^{-1} [h_d + h_p], \quad (10)$$

$$\text{var}(\text{vec}(\varphi) | \alpha, \Gamma, \delta, \Sigma \text{ data}) = [Q_d + Q_p]^{-1},$$

where:

$$Q_d = (\alpha' \Sigma^{-1} \alpha) \otimes (Z_2' Z_2), \quad Q_p = \omega_\varphi I_{sr}, \quad h_d = \text{vec}(Z_2' W \Sigma^{-1} \alpha), \quad h_p = \omega_\varphi \mu_\varphi$$

$$W = \Delta Y - Z_1 \alpha' - Y' \Gamma - D \delta, \quad Y_{.1} = [Z_1 | Z_2], \quad \text{with } Z_1 \text{ a } (p \times r) \text{ matrix.}$$

Proof: from the joint posterior pdf (9), the conditional posterior pdf of  $\varphi$  can be obtained as:

$$p(\varphi | \alpha, \Gamma, \delta, \Sigma \text{ data}) \propto \exp \{ - .5 [\text{trace} (E' E \Sigma^{-1} + \omega_\varphi (\text{vec}(\varphi) - \mu_\varphi)' (\text{vec}(\varphi) - \mu_\varphi))] \}.$$

Straightforward algebra leads then to the required result. ■

The justification of this result is straightforward: the combination of multivariate normal data evidence for  $\text{vec}(\varphi)$  and of multivariate normal prior information on it

generates a normal conditional posterior distribution whose moments are functions of data and prior moments.

A special case attains when the prior precision parameter is zero, i.e. in the absence of any a priori information. In this case, the conditional moments are:

$$E(\text{vec}(\Phi) | \alpha, \Gamma, \delta, \Sigma, \text{data}) = [Q_d]^{-1} h_d = \text{vec}[(Z_2' Z_2)^{-1} Z_2' W \Sigma^{-1} \alpha (\alpha' \Sigma^{-1} \alpha)^{-1}], \quad (11)$$

$$\text{var}(\text{vec}(\Phi) | \alpha, \Gamma, \delta, \Sigma, \text{data}) = [Q_d]^{-1} = (\alpha' \Sigma^{-1} \alpha)^{-1} \otimes (Z_2' Z_2)^{-1},$$

Notice that in this case, when  $\omega_\Phi$  is equal to zero, the precision matrix of the conditional posterior pdf of  $\Phi$  is not invertible when  $\alpha$  has less than full rank. When  $\omega_\Phi \neq 0$ , this case never arises.

The result contained in this lemma is important because it characterizes the marginal posterior distribution of the cointegration parameters contained in the matrix  $\Phi$ . In fact, it is immediate to notice that:

$$p(\text{vec}(\Phi) | \text{data}) = \int \int \int p(\Phi | \alpha, \Gamma, \delta, \Sigma, \text{data}) p(\alpha | \Gamma, \delta, \Sigma | \text{data}) d\alpha \, d\Gamma \, d\delta \, d\Sigma \quad (12)$$

What this expression says is that the finite sample marginal posterior distribution of  $\Phi$  is the average of multivariate normal distributions, weighted by the marginal distribution of the other parameters of the system. In other words they are mixtures of normals. From the viewpoint of the classical inference literature, this result has been shown by Johansen (1991, 1995a) to hold for the asymptotic distributions, and Phillips (1994) shows that the reduced rank regression cointegrating vectors have finite sample distributions with Cauchy tails and no moments. In this Bayesian

framework, the behaviour of  $p(\phi|data)$  does not present Cauchy tails when a proper prior for  $\phi$  is specified.

The conditional posterior distribution of  $\phi$  can be easily simulated.

**Lemma 4.2:** The conditional posterior distribution of  $\theta_2 = [vec(\alpha)' | vec(\Gamma)' | vec(\delta)']'$ , is  $(r \times s + p^2 \times (k-1) + d \times p)$ -variate normal, with moments:

$$E(\theta_2 | \phi, \Sigma, data) = [R_d + R_p]^{-1} [g_d + g_p],$$

$$var(\theta_2 | \phi, \Sigma, data) = [R_d + R_p]^{-1},$$

where:

$$R_d = S' [\Sigma^{-1} \otimes (X'X)] S, \quad R_p = \omega_\alpha I_{pr} + \Sigma_T^{-1} + \Sigma_\delta^{-1},$$

$$g_d = S' vec(X' \Delta Y \Sigma^{-1}), \quad g_p = \omega_\alpha \mu_\alpha + \Sigma_\delta^{-1} \mu_\delta, \quad X = [Y, \beta | Y' | D],$$

and  $S$  is a permutation matrix such that:

$$vec\{[\alpha | \Gamma | \delta]'\} = S \theta_2.$$

Proof: starting from (9), the conditional posterior pdf of  $\theta_2$  can be obtained as:

$$p(\theta_2 | \phi, \Sigma, data) \propto \exp\{-.5[trace(E'E \Sigma^{-1} + \omega_\alpha (vec(\alpha) - \mu_\alpha)' (vec(\alpha) - \mu_\alpha) + vec(\Gamma)' \Sigma_T^{-1} vec(\Gamma) + (\delta - \mu_\delta)' \Sigma_\delta^{-1} (\delta - \mu_\delta))]\}.$$

Some algebra leads then to the required result. ■

The intuition of this result is exactly in the same terms as in Lemma 4.1. On the basis of this result and of the properties of the multivariate Gaussian distribution, it is immediate to see that the conditional distributions  $p(\text{vec}(\alpha')|\Gamma, \delta, \phi, \Sigma, \text{data})$ ,  $p(\text{vec}(\Gamma)|\alpha, \delta, \phi, \Sigma, \text{data})$ ,  $p(\text{vec}(\delta)|\alpha, \Gamma, \phi, \Sigma, \text{data})$  are likewise normal.

A special case attains in the absence of any a priori information about  $\theta_2$ , i.e. when  $\omega_\alpha = 0$ ,  $\Sigma_1^{-1} = [0]$ ,  $\Sigma_\delta^{-1} = [0]$ . In this case, the conditional moments are:

$$E(\theta_2|\phi, \Sigma, \text{data}) = R_d^{-1}g_d = S^{-1}\{\text{vec}[(X'X)^{-1}X'\Delta Y]\},$$

$$\text{var}(\theta_2|\phi, \Sigma, \text{data}) = S^{-1}[\Sigma \otimes (X'X)^{-1}]S^{-1}.$$

The conditional posterior distribution of  $\theta_2$  can be easily simulated.

**Lemma 4.3:** The conditional posterior distribution of  $\Sigma$ ,  $p(\text{vech}(\Sigma)|\alpha, \phi, \Gamma, \delta, \text{data})$  is inverted Wishart.

Proof: starting from (9), the conditional posterior pdf of  $\Sigma$  is:

$$p(\Sigma|\alpha, \phi, \Gamma, \delta, \text{data}) \propto |\Sigma|^{-(T+n+1)/2} \exp\{-.5[\text{trace}(E'E\Sigma^{-1})]\}.$$

which can be recognised as inverted Wishart. ■

Such a distribution can be easily simulated: exploiting the properties of the Wishart distribution, one can easily draw from multivariate normal distributions, map this draw onto a draw from a Wishart distribution, and this latter one is mapped onto a draw for an inverted Wishart distribution, as required.

On the basis of these results, it is possible to generate as many draws from the marginal posterior pdf's as desired, and to put them in a Gibbs sampling sequence

which defines a Markov updating scheme. This scheme converges in distribution to the joint posterior pdf given that the conditions on the conditional pdf's described in Section [3.3c] for achieving convergence are satisfied.

Being able to generate draws on this distribution, it is possible to estimate the posterior expectation (if it exists) of any well defined function of the parameters, and the marginal posterior distributions of any subset of parameters of interest. These estimates are obtained on the basis of the Monte Carlo principle, to any desired degree of accuracy:

$$\bar{f}_N(\theta) = N^{-1} \sum_{i=1}^N f(\theta^{(i)}) \xrightarrow{a.s.} E(g(\theta|data)).$$

In order to obtain a Monte Carlo numerical estimate of the marginal posterior distribution of a certain subset of parameters, say  $\theta_1$ , the function  $f(\theta)$  is defined as  $p(\theta_1|\theta_2, \dots, \theta_k, data)$ :

$$p(\theta_1|data) \equiv N^{-1} \sum_{i=1}^N p(\theta_1|\theta_2^{(i)} \theta_3^{(i)} data),$$

whereas, in order to obtain the posterior moments of such distribution one could define  $f(\theta)$  as the corresponding conditional moment:

$$E(\theta_1|data) \equiv N^{-1} \sum_{i=1}^N E(\theta_1 | \theta_2^{(i)} \theta_3^{(i)} data),$$

$$var(\theta_1|data) \equiv N^{-1} \sum_{i=1}^N var(\theta_1 | \theta_2^{(i)} \theta_3^{(i)} data).$$

Due to the inherent correlation among draws in the Gibbs sample, the accuracy of the Monte Carlo estimates can be measured by means of heteroskedasticity-autocorrelation consistent (HAC) estimators of the standard error of the sample



mean of  $f(\theta)$ , based on a consistent estimate of its spectral density function at frequency zero. The simplest one, which delivers a well behaved estimate of the standard error, is the Newey and West estimator reviewed in Section [3.3c]. This estimator is used in the applications presented in this chapter. Following Geweke (1992), I also evaluated a *HAC* diagnostic test to assess whether convergence of the Gibbs Sampling scheme to the joint posterior distribution has occurred in the applications being presented in this chapter, testing the equality of the sample mean of a batch of early draws in the sequence and the sample mean of a batch of late draws in the sequence. Under the null of equality of the two sub-sample means, the resulting test statistic has an asymptotic standardised normal distribution. Acceptance of the null is interpreted as that the *GSS* has converged. For the details see Section [3.3.c].

#### [6.5] Inference on Cointegration Rank.

I now turn to the problem of how to conduct inference on the cointegration rank. The model described in the previous sections can be cast in a different parameterisation which is based on the singular value decomposition of  $\Pi = \alpha\beta'$  (see Dhrymes, 1978, p. 78):

$$\begin{aligned}\Pi &= U\Lambda V', \quad (\Pi\Pi')U = U\Lambda^2, \quad U'U = UU' = I_p, \\ (\Pi'\Pi)V &= V\Lambda^2, \quad V'V = VV' = I_p.\end{aligned}$$

The matrix  $\Lambda$  is diagonal with the square root of the eigenvalues of  $\Pi\Pi$ . Under the assumption of rank  $r < p$ , the singular value decomposition is:

$$\Pi = U_1\Lambda_1V_1', \quad U_1'U_1 = V_1'V_1 = I_r,$$

with  $U_1$  and  $V_1$  ( $p \times r$ ) matrices and  $\Lambda_1$  ( $r \times r$ ) diagonal matrix with the square roots of the positive eigenvalues of  $\Pi\Pi'$  on the diagonal. Thus the model can be equivalently written as:

$$\Gamma(L)\Delta y_t = U_1 \Lambda_1 V_1' y_{t-1} + \delta' D_t + \epsilon_t \quad (13)$$

Inference is then made on the number of diagonal elements of  $\Lambda_1$  being different from zero. The joint posterior distribution of the model as in expression (9) can be simulated by means of the Gibbs Sampling scheme described earlier. It is straightforward to map each draw on  $\alpha$  and  $\beta$  onto a draw on  $U_1$ ,  $\Lambda_1$  and  $V_1$  by applying the singular value decomposition to  $\Pi^{(i)} = \alpha^{(i)}\beta^{(i)'}.$  In this way it is possible to obtain a Monte Carlo estimate of the marginal posterior distribution of  $\lambda = \text{diag}(\Lambda_1)$  and of its moments, just by analytically characterizing the conditional posterior distribution of  $\lambda$ .

This is done in the following lemma.

**Lemma [5.1]:** The conditional posterior distribution of  $\lambda = \text{diag}(\Lambda_1)$  has the following kernel:

$$p(\lambda|U_1, V_1, \Gamma, \delta, \Sigma, \text{data}) \propto |\Omega|^{-1/2} \exp\{-0.5[(\lambda - \eta)' \Omega^{-1} (\lambda - \eta)]\},$$

$$\eta = c_1 - Q_{12} Q_{22}^{-1} c_2, \quad \Omega = Q_{11} - Q_{12} Q_{22}^{-1} Q_{21},$$

where:

$$\mathbf{c} = \begin{bmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \end{bmatrix} = \mathbf{G} \left[ \omega_{\alpha} (\mathbf{V}_{11}' \mathbf{V}_{11}) \otimes \mathbf{I}_r + (\mathbf{U}_1' \Sigma^{-1} \mathbf{U}_1) \otimes (\mathbf{V}_1' \mathbf{Y}_{-1}' \mathbf{Y}_{-1} \mathbf{V}_1) \right]^{-1} \times \\ \text{vec} \left[ \omega_{\alpha} \mathbf{U}_1' \mu_{\alpha} \mathbf{V}_{11} + \mathbf{V}_1' \mathbf{Y}_{-1}' \mathbf{W} \Sigma^{-1} \mathbf{U}_1 \right],$$

$$\mathbf{Q} = \begin{bmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12} \\ \mathbf{Q}_{21} & \mathbf{Q}_{22} \end{bmatrix} = \mathbf{G} \left[ \omega_{\alpha} (\mathbf{V}_{11}' \mathbf{V}_{11}) \otimes \mathbf{I}_r + (\mathbf{U}_1' \Sigma^{-1} \mathbf{U}_1) \otimes (\mathbf{V}_1' \mathbf{Y}_{-1}' \mathbf{Y}_{-1} \mathbf{V}_1) \right]^{-1} \mathbf{G}',$$

$$\mathbf{W} = \Delta \mathbf{Y} - \mathbf{Y}' \Gamma - \mathbf{D} \delta, \quad \mathbf{V} = \begin{bmatrix} \mathbf{V}_{11} \\ \mathbf{V}_{21} \end{bmatrix},$$

and  $\mathbf{G}$  is a permutation matrix such that  $\text{diag}(\Lambda_1)$  is given by the first  $r$  rows of  $\mathbf{G} \text{vec}(\Lambda_1)$ .

Proof: considering the parameterisation (13), the conditional posterior of  $\Lambda_1$  can be obtained from (9) as:

$$p(\Lambda_1 | \mathbf{U}_1, \mathbf{V}_1, \Gamma, \delta \text{ data}) \propto \exp \{ -0.5 [\text{trace} (\mathbf{E}' \mathbf{E} \Sigma^{-1} + \\ (\text{vec}(\mathbf{U}_1 \Lambda_1 \mathbf{V}_{11}') - \mu_{\omega})' \Sigma_{\omega}^{-1} (\text{vec}(\mathbf{U}_1 \Lambda_1 \mathbf{V}_{11}') - \mu_{\omega}))] \}.$$

Usual algebra gives the joint posterior of  $\Lambda_1$ , and applying the standard factorisation results for a multivariate Normal proves the lemma. ■

The conditional pdf's of the single elements of  $\text{diag}(\Lambda_1)$  are obtained by taking into consideration their nature as truncated normal distributions. For instance, the conditional pdf of the second element of  $\lambda$ ,  $\lambda_2$ , has support  $\lambda_2 \in (\lambda_3, \lambda_1)$ , and can be written as:

$$p(\lambda_2 | \lambda_1, \lambda_3, \mathbf{U}_1, \mathbf{V}_1, \Gamma, \delta \text{ data}) = \phi(\lambda_2 | \lambda_1, \lambda_3, \mathbf{U}_1, \mathbf{V}_1, \Gamma, \delta \text{ data}) / \\ [\Phi(\lambda_1 | \lambda_1, \lambda_3, \mathbf{U}_1, \mathbf{V}_1, \Gamma, \delta \text{ data}) - \Phi(\lambda_2 | \lambda_1, \lambda_3, \mathbf{U}_1, \mathbf{V}_1, \Gamma, \delta \text{ data})],$$

where  $\phi(\lambda_2 | \lambda_1, \lambda_3, U_1, V_1, \Gamma, \delta, \text{data})$  is a Gaussian pdf conditioned on,  $\lambda_1, \lambda_3, U_1, V_1, \Gamma, \delta$ , and  $\Phi()$  is the corresponding cdf.

On the basis of this analytical result, which holds for whichever rank of  $\Pi$ , from one to  $p$ , it is possible to conduct inference on the true cointegration rank, by means of the posterior distribution of  $\lambda = \text{diag}(\Lambda_1)$ . In the present context, rank equal to  $r$  is the maintained hypothesis. In order to check whether the rank is equal to  $r-1$ , one has to evaluate the posterior distribution of the  $r^{\text{th}}$  element of  $\lambda$  and see if zero falls within the highest posterior density confidence interval at a chosen confidence level (say 95%). This test has Johansen's  $\lambda$ -max test as a classical inference counterpart. In order to see whether it is possible to reduce the rank from  $r$  to  $r-2$ , one has to examine the joint posterior distribution of the last two elements of  $\lambda$ , and when the test is carried out at  $r = p$ , this has Johansen's *trace* test as a classical inference counterpart.

#### [6.6] Testing Restrictions on the Cointegration Space

Once the cointegrating rank has been decided, it might be interesting to check restrictions on the free parameters in the cointegrating vectors. We have already seen the lack of identification problem that has to be solved by imposing a certain structure on the  $\beta$  matrix. We choose to impose the normalisation  $\beta = [I, \phi']'$ . Remember that this structure does not impose any restriction on the space spanned by the cointegrating vectors. It is possible to impose restrictions on the cointegrating space, i.e. "overidentifying" restrictions on the columns of  $\beta$  of the kind:

$$\beta_i = H_i \phi_i, \quad i = 1, \dots, r.$$

The validity of these constraints can be tested by means of asymptotically  $\chi^2$  distributed *LR* statistics (Johansen, 1995b).

As for the finite sample performances of these tests, no analytical result is available. Recently, Cappuccio and Lubian (1995) have shown via Monte Carlo simulation that the empirical size of those tests is dramatically different from the nominal one, leading to systematic over-rejection of the maintained hypothesis also in fairly large sample sizes. For this reason, it is interesting to see what indications could be gathered by the use of Bayesian techniques based on finite sample evidence. Writing the over-identifying restrictions in the following form:

$$\mathbf{R}'\text{vec}(\boldsymbol{\varphi}) = \mathbf{d}, \quad (14)$$

I define the variable  $\boldsymbol{\xi} = \mathbf{R}'\text{vec}(\boldsymbol{\varphi}) - \mathbf{d}$ , whose conditional posterior distribution can be readily obtained from lemma 4.1 as *rs*-dimensional Normal with moments:

$$E(\boldsymbol{\xi} | \boldsymbol{\alpha}, \Gamma, \boldsymbol{\delta}, \Sigma \text{ data}) = \mathbf{R}'[\mathbf{Q}_d + \mathbf{Q}_p]^{-1}[\mathbf{h}_d + \mathbf{h}_p] - \mathbf{d},$$

$$\text{var}(\boldsymbol{\xi} | \boldsymbol{\alpha}, \Gamma, \boldsymbol{\delta}, \Sigma \text{ data}) = \mathbf{R}'[\mathbf{Q}_d + \mathbf{Q}_p]^{-1} \mathbf{R}.$$

If (14) holds, one would expect  $\boldsymbol{\xi}$  to have posterior pdf with expected value equal to zero. Defining  $SS = \boldsymbol{\xi}'\boldsymbol{\xi}$ , it is therefore possible to write :

$$E(SS | \text{data}) = \zeta = \text{trace}[\text{var}(\boldsymbol{\xi} | \text{data})].$$

Hence, on the basis of a Gibbs sample from the joint posterior distribution of  $\boldsymbol{\alpha}, \Gamma, \boldsymbol{\delta}, \Sigma$ , one could at each pass evaluate  $(SS)^{(i)}$ , and  $\text{var}(\boldsymbol{\xi} | \boldsymbol{\alpha}^{(i)}, \Gamma^{(i)}, \boldsymbol{\delta}^{(i)}, \Sigma^{(i)} \text{ data})$ ,  $i = 1, 2, \dots, N$ . This would allow one to obtain the posterior pseudo pdf of *SS* and:

$$\bar{\zeta}_N = N^{-1} \sum_{i=1}^N \text{trace}[\text{var}(\xi | \alpha^{(i)}, \Gamma^{(i)}, \delta^{(i)}, \Sigma^{(i)} | \text{data})],$$

a Monte Carlo consistent estimate  $\bar{\zeta}_N$  of  $\zeta$ . At this point, it is suggested to accept the hypothesis (14) at a desired confidence level, if the corresponding *HPD* for *SS* contains the value  $\bar{\zeta}_N$ .

Another testing strategy could be to evaluate the "LM" statistic at each pass of the *GSA* as:

$$LM^{(i)} = \xi^{(i)'} [\text{var}(\xi | \alpha^{(i)}, \Gamma^{(i)}, \delta^{(i)}, \Sigma^{(i)} | \text{data})]^{-1} \xi^{(i)}. \quad (15)$$

Also, measuring the distance of  $\xi$  from zero with a different metric, one could simulate the "LR" test at each step in two different ways:

$$\begin{aligned} LR_1^{(i)} &= T \ln \left[ \left| \Sigma_R^{(i)} \right| / \left| \Sigma^{(i)} \right| \right], \\ LR_2^{(i)} &= T \ln \left[ \left| \hat{\Sigma}_R^{(i)} \right| / \left| \hat{\Sigma}_{UR}^{(i)} \right| \right], \end{aligned} \quad (16)$$

where  $\Sigma^{(i)}$  is the *i*-th draw from  $p(\Sigma | \alpha^{(i)}, \varphi^{(i)}, \Gamma^{(i)}, \delta^{(i)}, \text{data})$ ,  $\Sigma_R^{(i)}$  is the *i*-th draw from  $p(\Sigma | \alpha^{(i)}, \mathbf{R}'\varphi^{(i)} = \mathbf{d}, \Gamma^{(i)}, \delta^{(i)}, \text{data})$ ,  $\hat{\Sigma}_{UR}^{(i)}$  is the *ML* estimate of  $\Sigma$  conditional on  $\alpha^{(i)}, \varphi^{(i)}, \Gamma^{(i)}, \delta^{(i)}$ , and  $\hat{\Sigma}_R^{(i)}$  is the *ML* estimate of  $\Sigma$  conditional on  $\alpha^{(i)}, \mathbf{R}'\varphi^{(i)} = \mathbf{d}, \Gamma^{(i)}, \delta^{(i)}$ .

The desired level *HPD* confidence intervals could be evaluated for these three statistics, and one could then check whether the value of *q*, which is the number of overidentifying restrictions actually being imposed falls within it or not. Notice that

the validity of the procedure is only asymptotical for the  $LR_1$  and  $LR_2$  statistics, which are intended to provide only additional corroborating evidence to the tests based on the finite sample posterior distributions of  $\xi\xi$  and  $LM$ .

### [6.7] Some Applications

In this section I present the results of four different applications of the technique described in the previous sections. The first application presented in this section is on a vector of simulated data. The main rationale behind this exercise is to gather information on how the procedure works, and on how the results obtained are precise, given perfect knowledge of the data generation process (*DGP*). The second and the third applications are on the Danish and Finnish money demand applications analysed by Johansen and Juselius (1990). The fourth application is on the *PPP-UIP* data for the UK studied in Johansen and Juselius (1992).

For all the applications I present the results of the base case of complete ignorance priors ( $\omega_\alpha = \omega_1 = 0$ ,  $\Sigma_1^{-1} = [0]$ ). As for hyperparameter  $\omega_\phi$ , setting it to a value different from zero will surely avoid local non-identification of  $\phi$ , which would occur all the times  $\alpha$  has deficient rank. For this reason, in all the three applications on "real" data I implement different values for this hyperparameter, and I monitor the sensitivity of the results in this respect.

In all the applications, the marginal posterior pdf's are obtained, when possible, via Monte Carlo integration of the corresponding conditional distributions, when the latter ones can be analytically computed. In the remaining cases, i.e. for the over-identifying restriction test statistics, the marginal posterior pdf's have been obtained by using the Gaussian kernel method with plug-in bandwidth (see Silverman, 1986).

For all the applications described in this section, the Monte Carlo simulations have been carried out on the basis of a sample of 10,000 Gibbs sampling draws, after having discarded the first 500 passes. A Bartlett window with bandwidth equal to 9 has been implemented to obtain the standard errors of the Monte Carlo estimates.

#### [6.7.1] A simulated data set example

The data generation process being used in the analysis is a very simple one:

$$\Delta y_t = \alpha \beta' y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N(0, \Sigma), \quad \alpha = [-0.3, -0.03], \\ \beta = [1, -1]', \quad \Sigma = \text{diag}\{0.01, 0.01\},$$

where the sample size  $T$  is equal to 200,  $y_t$  is obviously a bivariate  $I(1)$  process with zero mean differences. This is the simplest possible framework, given the low dimensionality of the model, the total absence of short run dynamics and of deterministic components. The model being estimated is:

$$\Delta y_t = \alpha \beta' y_{t-1} + \delta' D_t + \Gamma_1 \Delta y_{t-1} + \varepsilon_t,$$

where  $D_t$  is a deterministic vector containing a time trend and a set of 4 seasonal dummies. The "true" parameters  $\delta$  and  $\Gamma_1$  are therefore zero in the simulated data generation process.

The results are presented in Table 6.1 below. The individual parameters referred to in the table are defined as follows:  $\alpha = [\alpha_1 \ \alpha_2]'$ ,  $\beta = [1, \beta_2]'$ ,  $\lambda$  is the positive diagonal element of  $\Lambda$  under the assumption of cointegration rank one,  $\sigma_{11}$ ,  $\sigma_{12}$  and  $\sigma_{22}$  are the distinct element of  $\Sigma$ ,  $\gamma_{11}$ ,  $\gamma_{12}$ ,  $\gamma_{21}$ ,  $\gamma_{22}$  are the elements of  $\Gamma$ , and the



parameters associated with the deterministic component (four seasonal dummies

and a linear trend) are contained in  $\delta' = \begin{bmatrix} \psi_{11} & \psi_{12} & \psi_{13} & \psi_{14} & \psi_{15} \\ \psi_{21} & \psi_{22} & \psi_{23} & \psi_{24} & \psi_{25} \end{bmatrix}$ .

**Table 6.1: Simulation results**

	true value	post. mean est.	HAC std. error	conv. diagn.
$\alpha_1$	-0.3	-0.3225	0.0013	0.3027
$\alpha_2$	0.03	-0.0013	0.0013	0.5065
$\beta_2$	-1.00	-0.9881	--	--
$\lambda$	0.43	0.4579	0.0019	-0.2557
$\sigma_{11}$	0.01	0.0110	0.00003	0.8479
$\sigma_{12}$	0.00	0.0008	0.00002	0.0041
$\sigma_{22}$	0.01	0.0108	0.00003	-0.0564
$\gamma_{11}$	0.00	-0.0460	0.0018	-0.0623
$\gamma_{12}$	0.00	0.0009	0.0025	-0.4461
$\gamma_{21}$	0.00	-0.0329	0.0018	0.0234
$\gamma_{22}$	0.00	0.0305	0.0024	-0.0416
$\psi_{11}$	0.00	0.0292	0.0007	-0.1802
$\psi_{12}$	0.00	0.0038	0.0007	-0.1697
$\psi_{13}$	0.00	-0.0134	0.0007	-0.5543
$\psi_{14}$	0.00	-0.0129	0.0007	-0.2738
$\psi_{15}$	0.00	0.00008	0.000005	0.3754
$\psi_{21}$	0.00	-0.0182	0.0005	0.4801
$\psi_{22}$	0.00	-0.0151	0.0005	0.1604
$\psi_{23}$	0.00	-0.0111	0.0005	0.3823
$\psi_{24}$	0.00	-0.0067	0.0005	0.3534
$\psi_{25}$	0.00	0.00006	0.000003	-0.7380

Notes: The sample size being used is  $T=200$ . The hyperparameter  $\omega_p$  is set to zero, given that we surely do not have local identification problem here. The posterior means reported are obtained as sample averages over the draws. For the  $\beta_2$  parameter the mode of the posterior distribution is reported, given that the posterior expectation does not exist. The standard errors estimates are HAC in the Newey West specification with bandwidth =9 and Bartlett weights. The convergence diagnostics are obtained by comparing the results of the first 10% and the last 10% of the Gibbs sample of values for each one of the parameters.

Looking at the convergence diagnostic results, it is immediately noticeable that none of them is significant at the usual 5% size, and therefore it is possible to conclude that the Gibbs sampling scheme used in this analysis has reached convergence satisfactorily.

The results indicate that, even in the absence of prior information, it is possible to obtain quite precise information about the parameters of the model, in terms of their marginal posterior distribution and moments. The parameters in  $\alpha$  and  $\beta$  have posterior means very close to their true values. The *HAC* estimated standard errors are very small, notably the ones associated with the linear trend coefficients, whose rate of convergence is the fastest ( $T^{3/2}$ ). Together with the estimates of the posterior mean of  $\lambda$  as the Gibbs sample average of  $\lambda$  (0.4579) reported in the table, we present a further estimate of it in terms of:

$$E(\lambda|data) = N^{-1} \sum_{j=1}^N E(\lambda | \mathbf{U}_1^{(j)}, \mathbf{V}_1^{(j)}, \mathbf{\Gamma}^{(j)}, \boldsymbol{\delta}^{(j)}, \boldsymbol{\Sigma}^{(j)} | data) = 0.45225.$$

The posterior distributions have been obtained for  $\lambda$ , the main parameter of interest of the model, in terms of:

$$p(\lambda|data) = N^{-1} \sum_{j=1}^N p(\lambda | \mathbf{U}_1^{(j)}, \mathbf{V}_1^{(j)}, \mathbf{\Gamma}^{(j)}, \boldsymbol{\delta}^{(j)}, \boldsymbol{\Sigma}^{(j)} | data).$$

This marginal posterior distribution is the key element for conducting inference on the cointegrating rank. In fact it is possible to compare the rank one hypothesis with the rank zero hypothesis by constructing a highest posterior density confidence interval for  $\lambda$  and see whether  $\lambda=0$  falls inside or outside that interval. In the present context, the 95% *HPD*, obtained by means of numerical quadrature is [0.2, 0.7]. It is therefore uncontroversial that the hypothesis being supported by posterior evidence is rank equal to one.

#### [6.7.2] The Danish Money Demand Example

In this second application, following Johansen and Juselius (1990), I construct a

$VAR(2)$  model for the vector series  $y_t = [LRM, LRY, IB, ID]_t'$ , where  $LRM$  is the log of real  $M2$ ,  $LRY$  is the log of real income, and  $IB$  and  $ID$  are the logs of the gross bond and deposit interest rates respectively. The Danish quarterly data run from 1974:1 to 1987:3. The results are collected in Tables 6.2.1 and 6.2.2, and Figures 6.2.1 to 6.2.5 contain the posterior pdf's of the relevant parameters. In Figure 6.2.1 I present the univariate posterior pdf's of the parameters  $\lambda_1$  and  $\lambda_2$  obtained in a model where the cointegrating rank has been set equal to two. A weakly informative prior has been specified for the parameters in  $\beta$ , centered around the  $MLE$  estimate and with prior precision  $\omega_\phi = 0.5$ . Clearly, the second parameter has a posterior distribution with a large probability mass associated with values close to zero. As a consequence, the 95% highest posterior distribution confidence interval is  $[0.0, 0.58]$ , containing the value of zero. Therefore, I decide to work with a cointegrating rank equal to one.

Note that the  $MLA$  results contained in Johansen and Juselius (1990) do not give clear-cut support to the rank one hypothesis: conditional on the hypothesis  $\mu_0 = \alpha\beta_0$ , the *trace* test accepts rank equal to zero, whereas the  $\lambda$ -*max* test accepts rank equal to one, when using the customary 5% size. Exactly the same conclusions are reached by working with a model with an unrestricted intercept term. Using Johansen's (1992) sequential testing strategy described in Section [5.3], one would therefore end up by specifying a model with rank equal to zero and a restricted constant term.

Given the results of the Bayesian procedure for rank determination, I obtain the posterior distributions of the free parameters in  $\beta$ , i.e.  $\beta_{21}$ ,  $\beta_{31}$  and  $\beta_{41}$ . I tried different values for the hyperparameter  $\omega_\phi$  and in the figures 6.2.2 to 6.2.5 I present only the results obtained with  $\omega_\phi = 0$  (diffuse prior) and  $\omega_\phi = 0.05$  (very weakly informative prior). The prior distribution for  $\beta$  is centered around the

normalised reduced rank estimate conditional on  $r = 1$ . In Tables 6.2.1 and 6.2.2 only the results obtained with  $\omega_p = 0.05$  are presented, but sensitivity with respect to other choices of  $\omega_p$  has proved very small<sup>1</sup>.

The coefficient on *LRY* ( $\beta_{21}$ ) is centered around a modal value of -1.01, and the distributions of two coefficients on the interest rates are centered around modal values which have the expected signs. Following Johansen and Juselius (1990), I tried to verify two hypotheses. The first one is that of unit elasticity of money with respect to income, i.e. that  $\beta_{21} = -1$ . In the classical inferential setting, this hypothesis can be represented as:

$$\beta = H\phi, H = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \phi = \begin{bmatrix} \phi_{11} \\ \phi_{21} \\ \phi_{31} \end{bmatrix},$$

and tested by means of an asymptotically  $\chi^2$  distributed *LR* test.

The second hypothesis of interest is that the difference between the two interest rates measures the opportunity cost of holding money in the long run equilibrium relationship. In the classical inferential setting, this is accomplished by means of an asymptotically  $\chi^2$  distributed *LR* test of the hypothesis:

$$\beta = H\phi, H = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & -1 \end{bmatrix}, \phi = \begin{bmatrix} \phi_{11} \\ \phi_{21} \\ \phi_{31} \end{bmatrix}.$$

In Johansen and Juselius (1990), the asymptotic test leads to joint acceptance of the two hypotheses.

In the present Bayesian framework, one immediately sees that the hypothesis of unit income elasticity ( $\beta_{21} = -1$ ) is clearly supported by the data: looking at the

<sup>1</sup> In all tables concerning the *HPD* credible sets constructed for checking the over-identifying constraints, the results of the simulation of the  $LR_2$  statistic have not been reported, since they always practically coincide with the ones regarding  $LR_1$ .

marginal posterior pdf of  $\beta_{21}$  (see Figures 6.2.2.b and 6.2.4.b), the value of -1 is within the 95% HPD confidence interval.

The plausibility of the second hypothesis, i.e.  $\beta_{31} = -\beta_{41}$ , can be gauged on the basis of the posterior distribution of the relevant statistics  $LR_1$ ,  $LR_2$ ,  $SS$  and  $LM$  (see Figure 6.2.5 and Table 6.2.2). The posterior empirical distributions of  $LR_1$ ,  $LR_2$  and  $LM$  are very much concentrated away from  $q = 1$ , and the 95% HPD interval constructed on the posterior distribution of  $SS$  does not contain the estimate of  $\text{trace}[\text{var}(\xi|data)]$ . Therefore, unlike Johansen and Juselius (1990), I reject the hypothesis  $\beta_{31} + \beta_{41} = 0$  on the basis of finite sample evidence. This finding does not conflict with the view that the two interest rates are jointly related to some measure of the opportunity cost of holding money, but rather that this opportunity cost is not properly measured by the interest rate differential.

**Table 6.2.1.** Results from Danish money demand example (see Johansen and Juselius, 1990). The VAR has four variables ( $LRM$ ,  $LRY$ ,  $IB$ ,  $ID$ ), 2 lags, a constant and a set of three centered seasonal dummies.

	post. mean est.	HAC std. error	conv. diagn.
$\alpha_{11}$	-0.1669	0.0023	0.2004
$\alpha_{21}$	0.1098	0.0024	-0.0761
$\alpha_{31}$	0.0127	0.0008	-0.6163
$\alpha_{41}$	0.0223	0.0006	-0.5035
$\beta_{21}$	-1.0127	0.0077	0.1164
$\beta_{31}$	5.4540	0.0395	-0.3204
$\beta_{41}$	-4.5936	0.0699	0.4233
$\lambda_1$	1.5403	0.0118	-0.7249

Notes: The sample size is  $T = 53$ . The posterior means reported are obtained as sample averages over the draws. A Gibbs sample of 10,000 replications is drawn from  $p(\alpha, \beta, \Sigma|data)$ , once  $\delta$  and  $\Gamma$  have been marginalised out. Standard errors estimates are Newey-West with bandwidth = 9 and Bartlett weights. Convergence diagnostics are obtained by comparing the results of the first 10% and the last 10% of the Gibbs sample. The hyperparameter  $\omega_\phi$  is set equal to 0.05.

**Table 6.2.2** Danish money demand. Testing for  $\beta_{31} + \beta_{41} = 0$ .  
95% HPD credible sets.  
Model with  $\omega_m = 0.05$ .

HPD set for $LR_1$	[2.12, 19.55]
HPD set for $LM$	[40.08, 304.86]
HPD set for $SS$	[15.93, 122.84]
$trace[var(\xi data)]$	0.90

### [6.6.3] The Finnish Money Demand Example

In this third application, following Johansen and Juselius (1990), I construct a  $VAR(2)$  model for the vector series  $y_t = [MON \ IRATE \ INF \ INC]_t'$ , where  $MON$  is the log of real  $M_0$ ,  $INC$  is the log of real income,  $IRATE$  is the log of the Bank of Finland marginal rate, and  $INF$  is the inflation rate. The quarterly Finnish data run from 1958:2 to 1984:3. The results are collected in Tables 6.3.1 and 6.3.2, and Figures 6.3.1 to 6.3.6 contain the posterior pdf's of the relevant parameters. As in the previous application the prior pdf for  $\phi$  is centered around its  $MLE$  estimate, and the results are graphed for  $\omega_\phi = 0.0$  and  $\omega_\phi = 0.1$ . In Figure 6.3.1 I present the univariate posterior pdf's of the parameters  $\lambda_1$ ,  $\lambda_2$ ,  $\lambda_3$ , obtained in a model where the cointegrating rank has been set equal to three. Since the 95% HPD confidence interval for  $\lambda_3$  does not contain the value of zero, I decided to work with a cointegrating rank equal to three. This coincides with the decision taken by Johansen and Juselius (1990) in this respect, but it is necessary to point out that their decision is based on the results of their  $ML$  asymptotic rank tests where they choose to work with a size of 20%. The finite sample Bayesian results seem to place this finding on a firmer ground. Given this assumption on the rank, the

normalized  $\beta$  becomes:  $\beta = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \beta_{41} & \beta_{42} & \beta_{43} \end{bmatrix}$

As for the structural interpretation of the long-run coefficients, Johansen and Juselius (1990) test and accept the hypothesis that in all cointegrating vectors the income elasticity is unity. In other words they test the hypothesis:

$$\beta = H\Phi, H = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & 0 & 0 \end{bmatrix}, \Phi = \begin{bmatrix} \phi_{11} & \phi_{12} & \phi_{13} \\ \phi_{21} & \phi_{22} & \phi_{23} \\ \phi_{31} & \phi_{32} & \phi_{33} \end{bmatrix}$$

This hypothesis amounts to stating that the log money/income ratio, the inflation and the interest rate series are all  $I(0)$  variables.

In the present Bayesian framework, we can check the hypothesis of stationarity of each series by inspection of the univariate posterior pdf's of the parameters  $\beta_{41}$ ,  $\beta_{42}$  and  $\beta_{43}$ , respectively (see Figures 6.3.2 and 6.3.5). The univariate 95% *HPD* intervals contain the values of -1, 0, and 0, and this circumstance favours the hypothesis of stationarity of each series.

It is possible to gauge the plausibility of the joint hypothesis  $\beta_{41}=-1$ ,  $\beta_{42}=\beta_{43}=0$  by inspection of the posterior distributions of the statistics  $LR_1$ ,  $LR_2$ ,  $SS$  and  $LM$  (see Figures 6.3.3 and 6.3.6 and Table 6.3.2). The 95% *HPD* confidence intervals for  $LR_1$ ,  $LR_2$ ,  $SS$  do contain the value  $q = 2$ , and the 95% *HPD* confidence interval for  $SS$  contain the estimate of  $trace[var(\xi|data)]$ . Therefore, all these results make a case in favour of accepting the hypothesis  $\beta_{41}=\beta_{42}=\beta_{43}=0$ .

In synthesis, as for the interpretation of the cointegrating coefficients, the results of the Bayesian testing procedure confirm the conclusions drawn from the application of the asymptotic *MLA* test.

It seems necessary to remark that the two money demand examples being considered in this chapter lead to completely different results. In the Danish money demand example a true long term equilibrium relationship is obtained, involving all the variables being considered in the *VAR* system. In the Finnish example things are radically different, since two of the series being considered are stationary and hence

they do not appear in the only meaningful long run equilibrium relationship given by the unit long run income elasticity of money demand.

**Table 6.3.1** Results from Finnish money demand example (see Johansen and Juselius, 1990). The *VAR* has four variables (*MON IRATE INF INC*), 2 lags, a constant and a set of three centered seasonal dummies.

	post mean est.	HAC std error	conv.diagn.
$\alpha_{11}$	-0.0852	0.0019	0.5122
$\alpha_{21}$	0.0918	0.0013	-0.1355
$\alpha_{31}$	0.0002	0.0004	0.1538
$\alpha_{41}$	0.0242	0.0011	0.2043
$\alpha_{12}$	-0.2182	0.0033	-0.3949
$\alpha_{22}$	-0.4613	0.0025	-0.4858
$\alpha_{32}$	-0.0022	0.0008	0.1522
$\alpha_{42}$	-0.1389	0.0020	-0.2154
$\alpha_{13}$	-0.4838	0.0161	-0.4311
$\alpha_{23}$	0.6457	0.0117	0.4797
$\alpha_{33}$	-0.4390	0.0038	-0.7555
$\alpha_{43}$	-0.2564	0.0098	-0.1062
$\beta_{41}$	-0.9556	0.0148	0.3226
$\beta_{42}$	0.0181	0.0026	0.4077
$\beta_{43}$	-0.0146	0.0007	-1.0889
$\lambda_1$	1.2359	0.0096	0.4488
$\lambda_2$	0.3080	0.0032	0.6122
$\lambda_3$	0.1543	0.0018	-0.4927

Notes: The sample size being used is  $T=62$ . The posterior means reported are obtained as sample averages over the draws. A Gibbs sample of 10,000 replications is drawn from the joint posterior  $p(\alpha \beta \Sigma | \text{data})$ , once the parameters in  $\delta$  and  $\Gamma$  have been marginalised out at the outset. The standard errors estimates are *HAC* in the Newey West specification with bandwidth =9 and Bartlett weights. The convergence diagnostics are obtained by comparing the results of the first 10% and the last 10% of the Gibbs sample. The hyperparameter  $\omega_0$  is set equal to 0.1.

**Table 6.3.2.** Finnish money demand. Testing for  $\beta_{41}=-1$ ,  $\beta_{42}=\beta_{43}=0$ .  
95% HPD credible sets.  
Model with  $\omega_0=0.1$ .

HPD set for $LR_1$	[2.28, 65.22]
HPD set for $LM$	[1.27, 32.47]
HPD set for $SS$	[0, 0.26]
$\text{trace}[\text{var}(\xi   \text{data})]$	0.24



#### [6.7.4] The UK PPP/UIP Example

In this fourth application, following Johansen and Juselius (1992), I construct a  $VAR(2)$  model for the vector series  $y_t = [P_1, I_1, P_2, I_2, E_{12}]'_t$ , where  $P_1$  is the log UK prices,  $P_2$  is the log trade weighted foreign price index,  $E_{12}$  is the log UK effective exchange rate,  $I_1$  is the log 3-month UK treasury bill rate and  $I_2$  is the log 3-month Eurodollar rate. The quarterly UK data runs from 1972:1 to 1987:2. Following Johansen and Juselius (1992), the  $VAR$  model has been augmented to include current and lagged values of  $DPOIL$ , the first differences in the log oil price series

Figures 6.4.1 to 6.4.11 contain the posterior pdfs of the relevant parameters. In Figure 6.4.1 I present the univariate posterior pdfs of the parameters  $\lambda_1, \lambda_2, \lambda_3$  obtained in a model where the cointegrating rank has been set equal to three, with a vaguely informative prior for  $\phi$  centered around the normalised  $MLA$  estimate. The third parameter has a posterior distribution with a large probability mass associated with values close to zero. As a consequence, the 95%  $HPD$  confidence interval contains the value of zero. The other three parameters have posterior distributions assigning negligible probabilities to the neighbourhood of zero. Therefore, I decided to work with a cointegrating rank equal to two. Again, this is a finite sample result that happens to coincide with the decision taken by Johansen and Juselius (1992) in this respect, but it is necessary to point out that their 5% size  $ML$  asymptotic rank test results do not allow them to do so. The finite sample Bayesian results again provide a firmer foundation to this conclusion.

In the favoured model with cointegrating rank equal to two, I then decided to work with a prior distribution for the free elements of  $\beta$  centered around the values implied by the validity of the  $PPP/UIP$  hypotheses.

I now turn to the interpretation of the cointegrating coefficients. Given this assumption on the rank, the normalized  $\beta$  becomes:

$$\beta = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \beta_{31} & \beta_{32} \\ \beta_{41} & \beta_{42} \\ \beta_{51} & \beta_{52} \end{bmatrix}$$

The hypothesis of interest is that the *PPP* and *UIP* relationships hold. The validity of the *PPP* clearly imposes the following three constraints on the normalised  $\beta$  matrix:

$$\beta_{31} = \beta_{51} = -1, \beta_{41} = 0,$$

in other words, the *PPP* imposes stationarity on the real exchange rate.

The validity of *UIP* hypothesis imposes non linear constraints on the autoregressive representation coefficients. Such constraints are the "hallmark" of rational expectation models; see Campbell and Shiller (1987) for an example in this respect. Following Johansen and Juselius (1992), I only focus on the implication that, if the *UIP* holds and the nominal exchange rate is  $I(1)$ , then the interest rate differential must be stationary. Note however that, strictly speaking, interest rate differential stationarity is implied by the *UIP* hypothesis but not viceversa.

This second set of constraints can therefore be written as follows:

$$\beta_{32} = \beta_{52} = 0, \beta_{42} = -1.$$

Taking all the restrictions into consideration, we end up then with a set of  $q=6$  constraints. In the classical inferential setting, inference is accomplished by means of an asymptotically  $\chi^2$  distributed LR test of the hypothesis:

$$\begin{aligned} \beta_1 &= H_1 \phi_1, \quad H_1 = \begin{bmatrix} 1 & 0 & -1 & 0 & -1 \end{bmatrix}', \quad \phi_1 = [\phi_{1,11}], \\ \beta_2 &= H_2 \phi_2, \quad H_2 = \begin{bmatrix} 0 & 1 & 0 & -1 & 0 \end{bmatrix}', \quad \phi_2 = [\phi_{2,11}], \end{aligned}$$

where  $\beta = [\beta_1 | \beta_2]$ .

In the present Bayesian framework, it is possible to gauge the plausibility of this hypothesis by inspection of the posterior distributions of the statistics  $LR_1$ ,  $LR_2$ ,  $SS$

and  $LM$  (see Figures 6.4.4 and 6.4.9 and the first column of Table 6.4.2). The 95% *HPD* confidence intervals for  $LR_1$ ,  $LR_2$ ,  $LM$  do not contain the value  $q = 6$ , and the 95% *HPD* confidence interval for  $SS$  does not contain 2.29, i.e. the estimate of  $trace [var (\xi | data)]$ . Therefore, all these result make a case against accepting the null hypothesis.

It is also possible to test separately the validity of the *PPP* and of the *UIP* hypotheses, by obtaining the relevant  $LR_1$ ,  $LR_2$ ,  $LM$  and  $SS$  statistics at each pass of the *GSS*. The corresponding empirical posterior distribution are graphed in Figures 6.4.5 and 6.4.10 for the *PPP* hypothesis, and in Figures 6.4.6 and 6.4.11 for the *UIP* hypothesis.

The results of this analysis clearly indicate that the *PPP* hypothesis fails to hold: looking at the second column of Table 6.4.2, the 95% *HPD* credible sets for  $LR_1$ ,  $LR_2$  and  $LM$  do not contain 3, and the corresponding interval for  $SS$  does not contain the estimate of  $trace [var (\xi | data)]$ .

As for the *UIP* hypothesis, looking at the third column of Table 6.4.2, the 95% *HPD* confidence intervals for  $LR_1$ ,  $LR_2$  and  $LM$  contain 3, and the corresponding interval for  $SS$  contain the estimate of  $trace [var (\xi | data)]$ . On the basis of *ML* asymptotic test statistics, Johansen and Juselius (1992) reach the same conclusions. Clearly, failure of the *PPP* to hold in the long run is not really appealing. A possible explanation to the non-stationarity of the real exchange rate can be related to the measurement of the price indices being analysed. As Johansen and Juselius (1992) point out, the two countries being considered could have experienced different productivity growths in the sample period, or they could have been characterised by differing proportions of tradeable goods. Another likely explanation is given by the fact that stationarity of the real exchange rate is a consequence of international arbitrage taking place in the goods markets. Such

arbitrage is costly and therefore can operate with considerable lags, in this way inducing the observation of non-stationary real exchange rates in finite samples.

**Table 6.4.1.** Results from the UK PPP/UIP example (see Johansen and Juselius, 1992). The VAR has four variables ( $P_1 I_1 P_2 I_2 E_{12}$ ), 2 lags, a constant and a set of three centered seasonal dummies, and the current and lagged values of *DPOIL* as exogenous variables

	post mean est	HAC std. error	conv. diagn.
$\alpha_{11}$	-0.0641	0.0008	0.2816
$\alpha_{21}$	0.0648	0.0013	0.0751
$\alpha_{31}$	-0.0168	0.0008	0.2837
$\alpha_{41}$	0.0114	0.0011	0.6601
$\alpha_{51}$	0.0054	0.0032	0.5555
$\alpha_{12}$	0.0873	0.0025	-0.2544
$\alpha_{22}$	-0.2274	0.0041	-0.1254
$\alpha_{32}$	0.0519	0.0024	-0.8593
$\alpha_{42}$	0.0284	0.0035	-0.8211
$\alpha_{52}$	-0.1533	0.0100	-0.2721
$\beta_{31}$	-1.3096	0.0076	0.4049
$\beta_{41}$	-1.8131	0.0387	0.2137
$\beta_{51}$	-0.3086	0.0152	0.8295
$\beta_{32}$	-0.0943	0.0033	-0.2611
$\beta_{42}$	-0.3106	0.0226	0.6154
$\beta_{52}$	0.1860	0.0063	0.8263
$\lambda_1$	0.6458	0.0061	-0.0366
$\lambda_2$	0.2402	0.0031	-0.7353

Notes: The sample size being used is  $T=63$ . The posterior means reported are obtained as sample averages over the draws. A Gibbs sample of 10,000 replications is drawn from the joint posterior  $p(\alpha \beta \Sigma | data)$ , once the parameters in  $\delta$  and  $\Gamma$  have been marginalised out at the outset. The standard errors estimates are HAC in the Newey West specification with bandwidth =9 and Bartlett weights. The convergence diagnostics are obtained by comparing the results of the first 10% and the last 10% of the Gibbs sample. The hyperparameter  $\omega_0$  is set equal to 0.1.

**Table 6.4.2.** UK exchange rate data. Testing for hypotheses: (a) stationarity of real exchange rate and interest rate differential. (b) Stationarity of real exchange rate. (c) Stationarity of interest rate differential. 95% *HPD* credible sets. Model with  $\omega_{\varphi}=0.1$ .

	Hypothesis (a)	Hypothesis (b)	Hypothesis (c)
<i>HPD</i> set for $LR_1$	[23.48, 89.21]	[28.01, 108.92]	[0, 93.79]
<i>HPD</i> set for $LM$	[28.53, 190.63]	[10.45, 120.78]	[1.74, 35.56]
<i>HPD</i> set for $SS$	[3.02, 48.02]	[2.08, 38.02]	[.02, 6.09]
$trace[var(E data)]$	2.29	1.87	0.42

### [6.8] Conclusion

Some brief general comments about the results of the application described in this chapter seem necessary.

Note that when an informative prior is used on the free elements of the cointegrating matrix, the posterior univariate pdfs of these parameters do not show the Cauchy-like tails presented in the case when the prior is diffuse. In order to realise this, it is just sufficient to compare, for instance, Figures 6.4.3 and 6.4.8. Imposing even a weakly informative prior has then the effect of trimming off these huge tails. In this way, it is possible to avoid the problem encountered using the maximum likelihood estimator of the cointegrating coefficients, whose finite sample properties are badly affected by these Cauchy-like tails.

Moreover, when a diffuse prior for  $\varphi$  is specified, the simulated posterior distribution of the  $LM$  statistic becomes very unstable. This does not happen when a proper prior for  $\varphi$  is used. This is clearly a consequence of the fact that in this latter case the posterior distributions of the coefficients of  $\varphi$  have finite variances.

Figure 6.2.1: Danish Money Demand

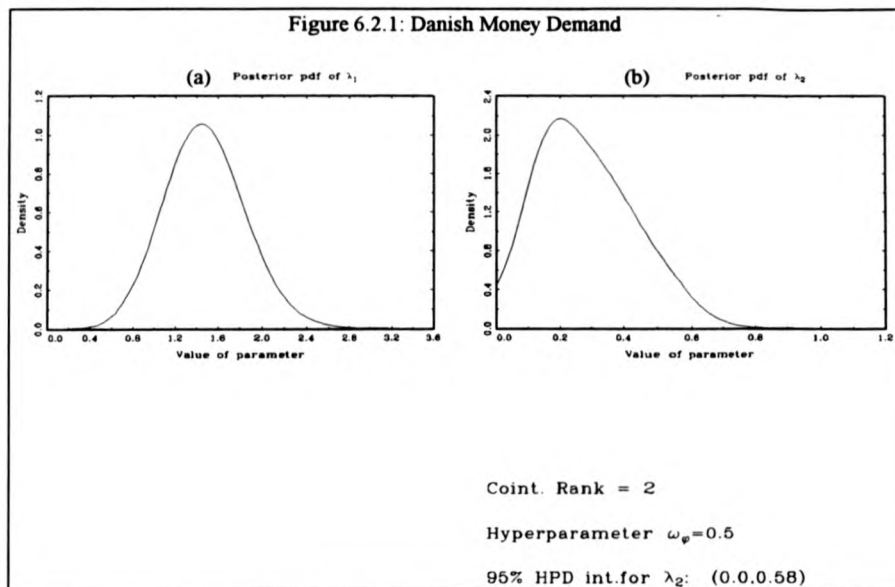


Figure (6.2.2): Danish Money Demand. Cointegrating rank = 1

Hyperparameter  $\omega_\phi = 0.0$

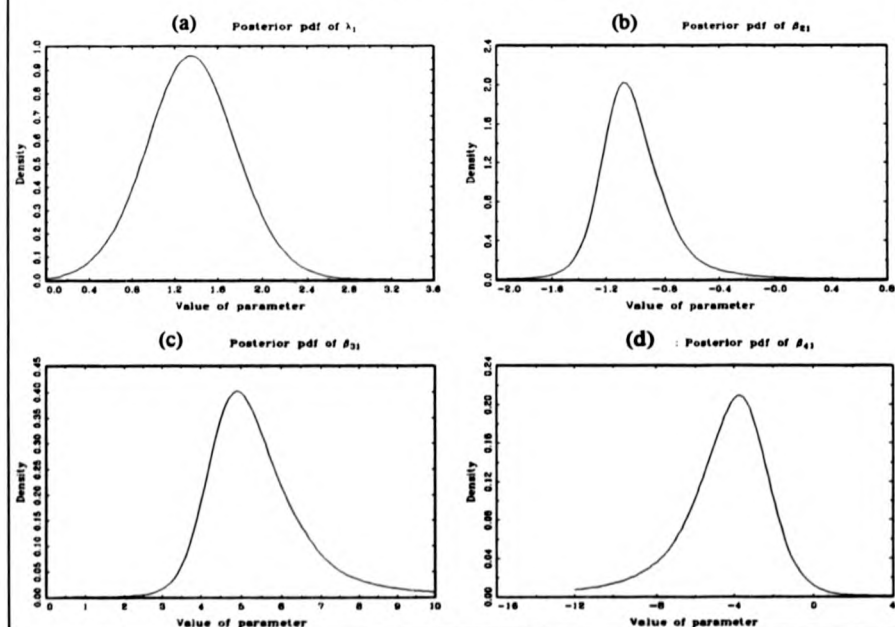


Figure (6.2.3): Danish Money Demand. Cointegrating rank =1  
Hyperparameter  $\omega_\psi = 0.0$

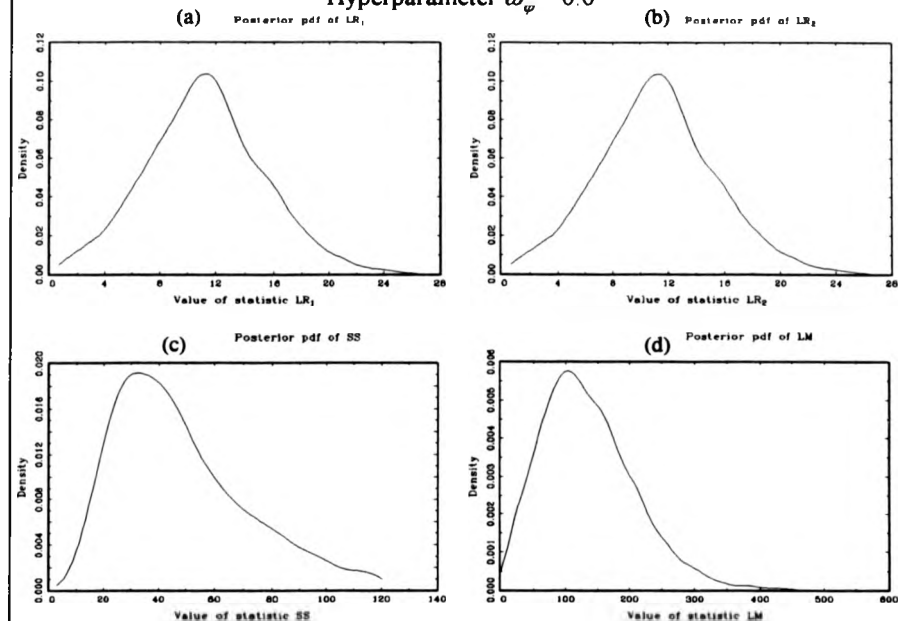
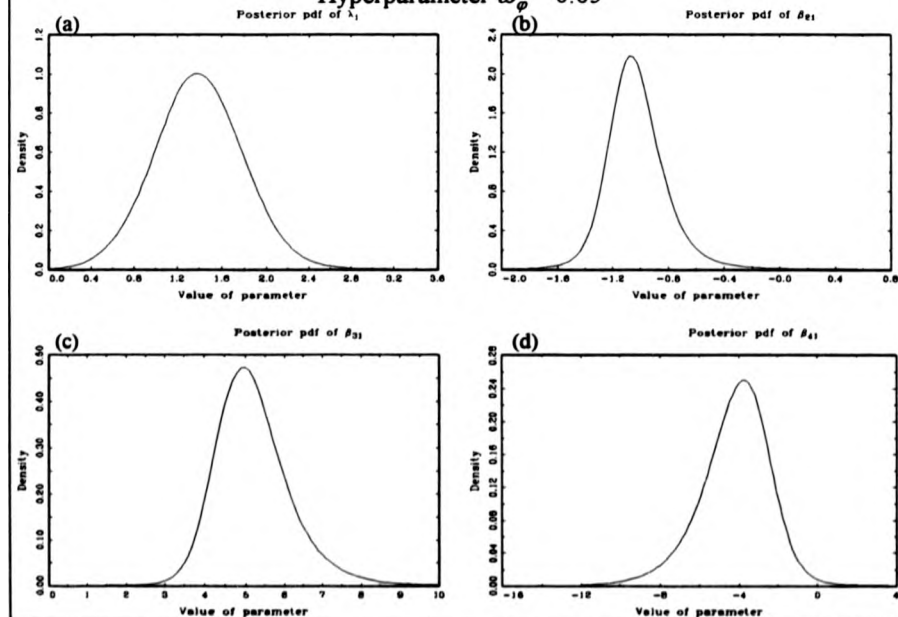


Figure (6.2.4): Danish Money Demand. Cointegrating rank =1  
Hyperparameter  $\omega_\psi = 0.05$



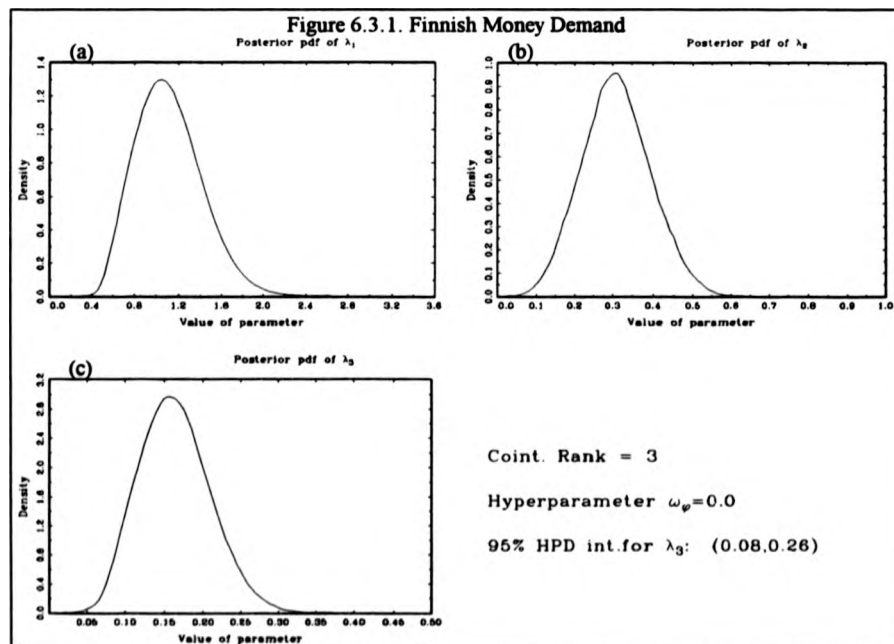
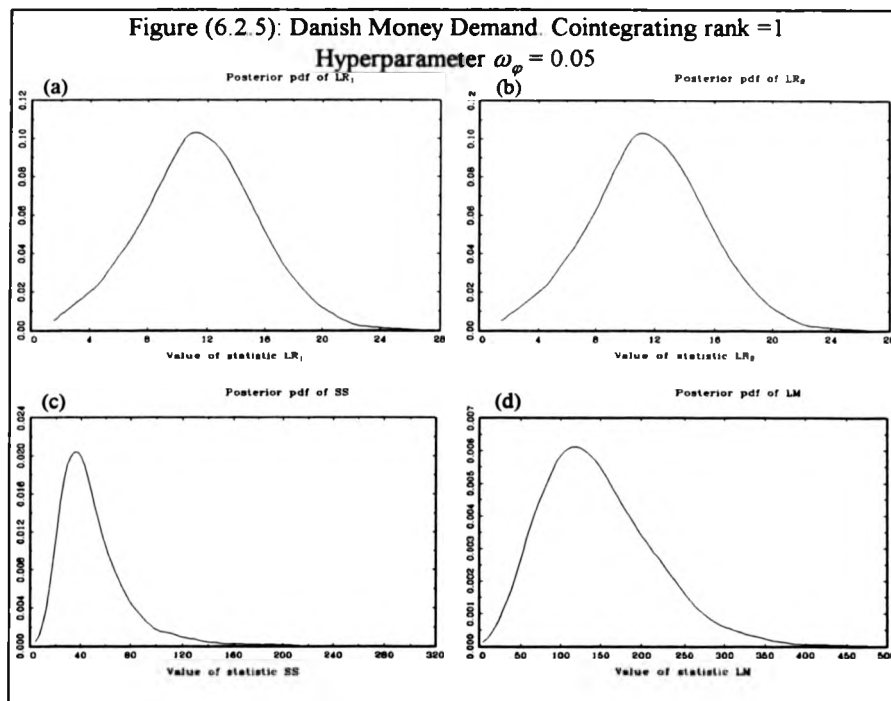




Figure 6.3 2. Finnish Money Demand

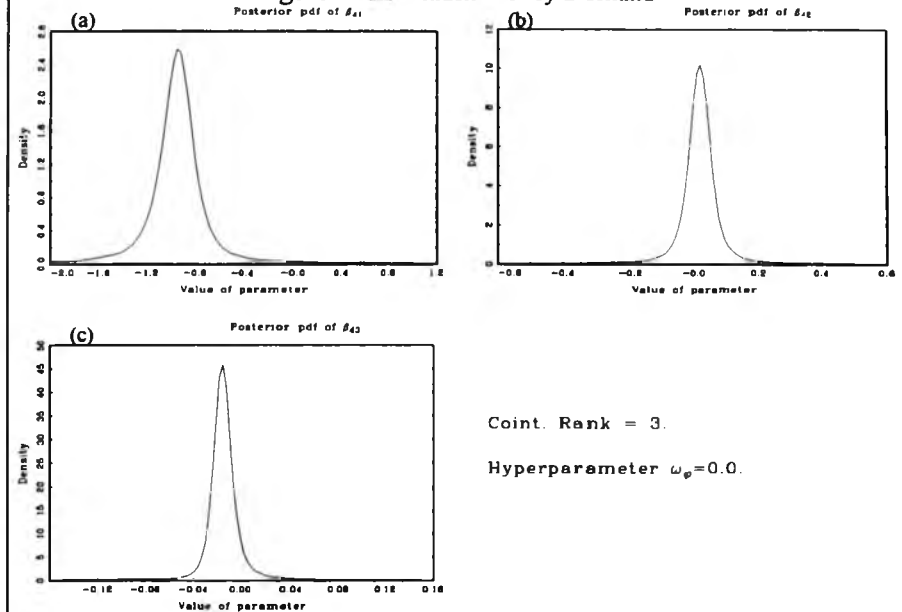


Figure (6.3.3). Finnish Money Demand. Cointegrating rank=3.  $\omega_\phi=0.0$

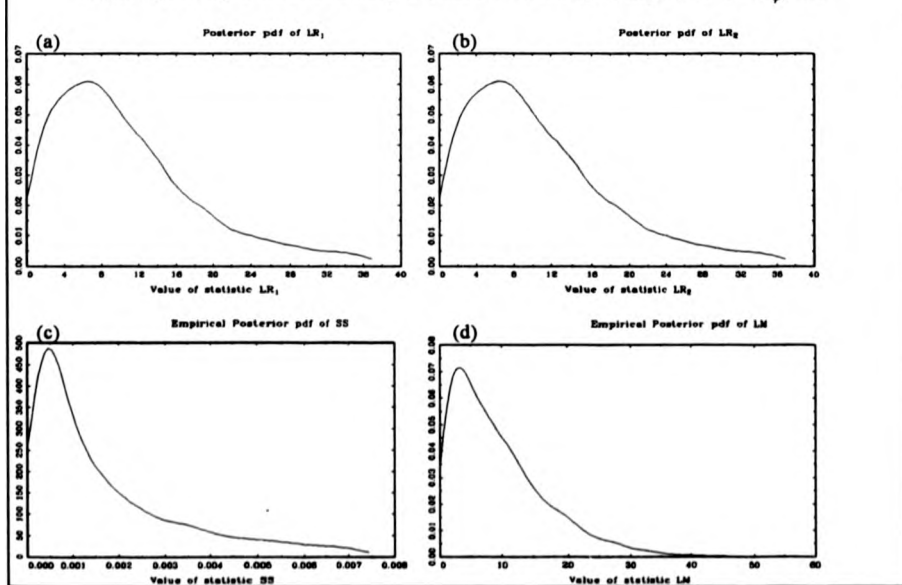


Figure 6.3.4. Finnish Money Demand

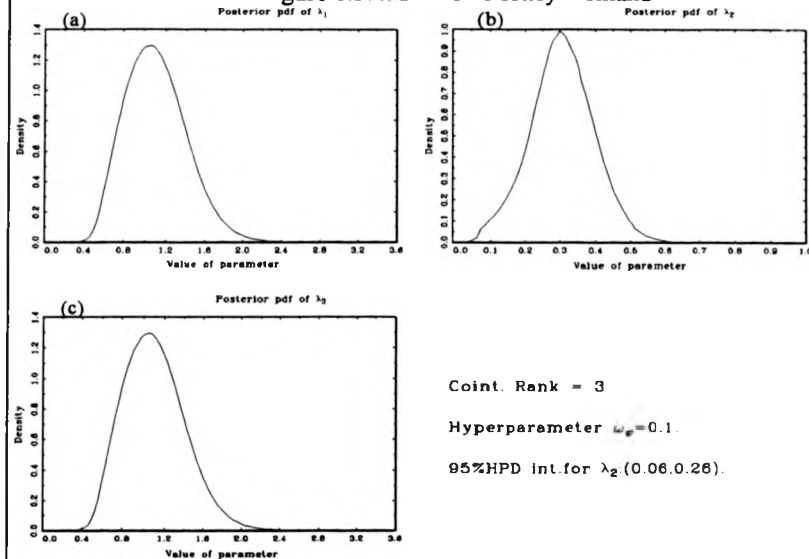


Figure 6.3.5. Finnish Money Demand

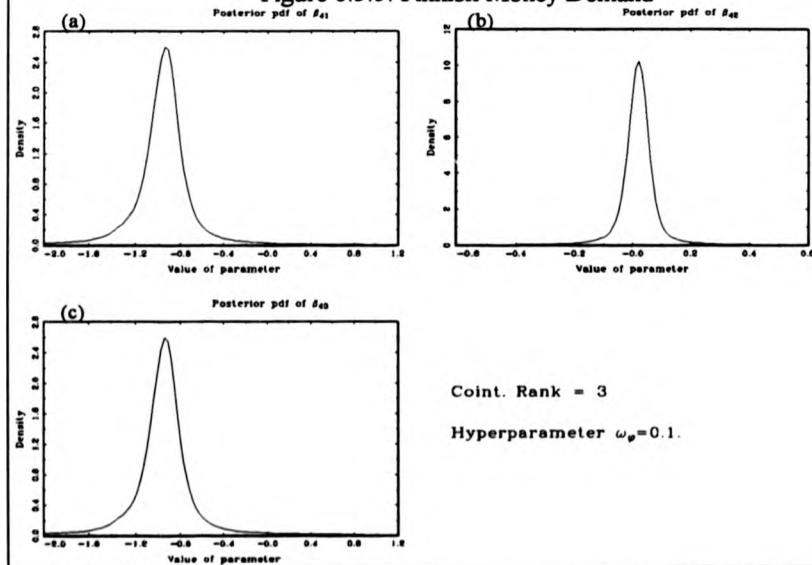


Figure (6.3.6). Finnish Money Demand. Cointegrating rank=3.  $\omega_\varphi=0.1$

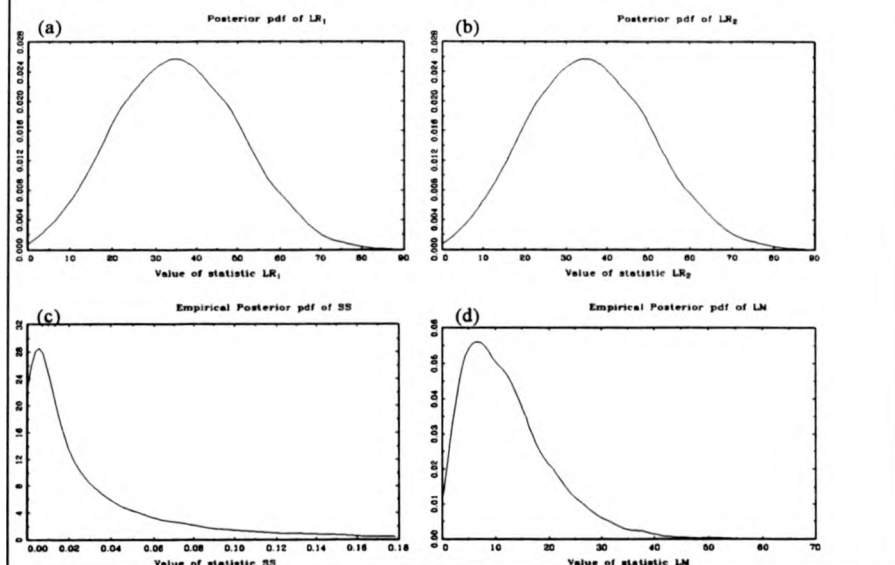


Figure 6.4.1. UK PPP/UIP data

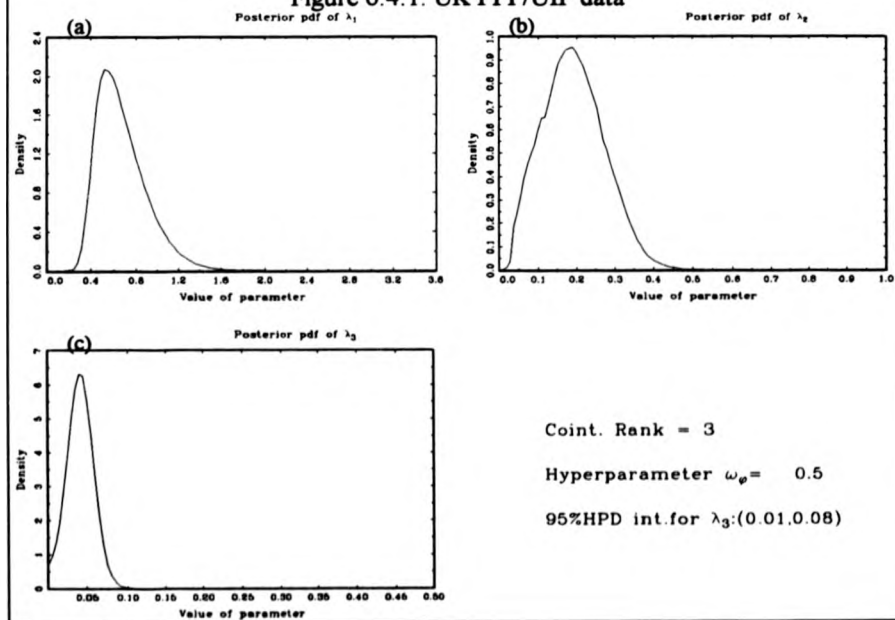


Figure (6.4.2). UK PPP/UIP. Coimt. rank=2.  $\omega_\phi=0.0$ . 95%HPD int. for  $\lambda_2$ : (0.02, 0.31)

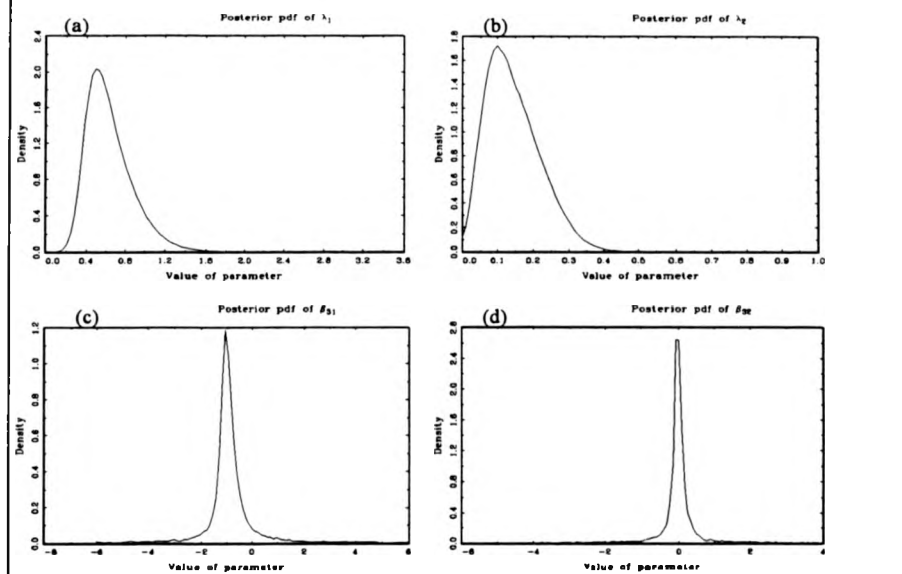


Figure (6.4.3). UK PPP/UIP. Coimt. rank=2.  $\omega_\phi=0.0$ .

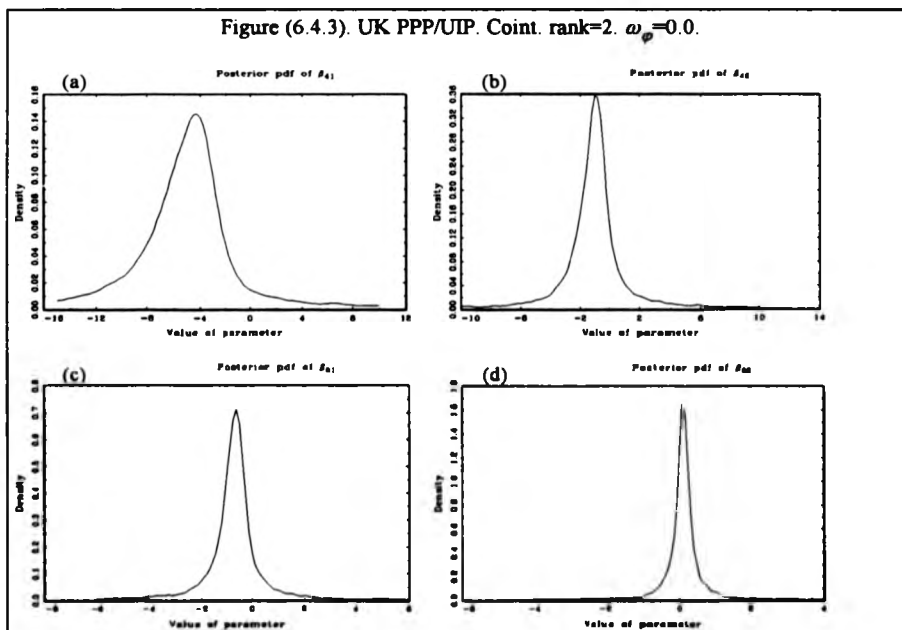


Figure (6.4.4). UK PPP/UIP. Co-int. rank=2.  $\omega_\phi=0.0$ .  
Testing stationarity of real exchange rate and int. rate differential.

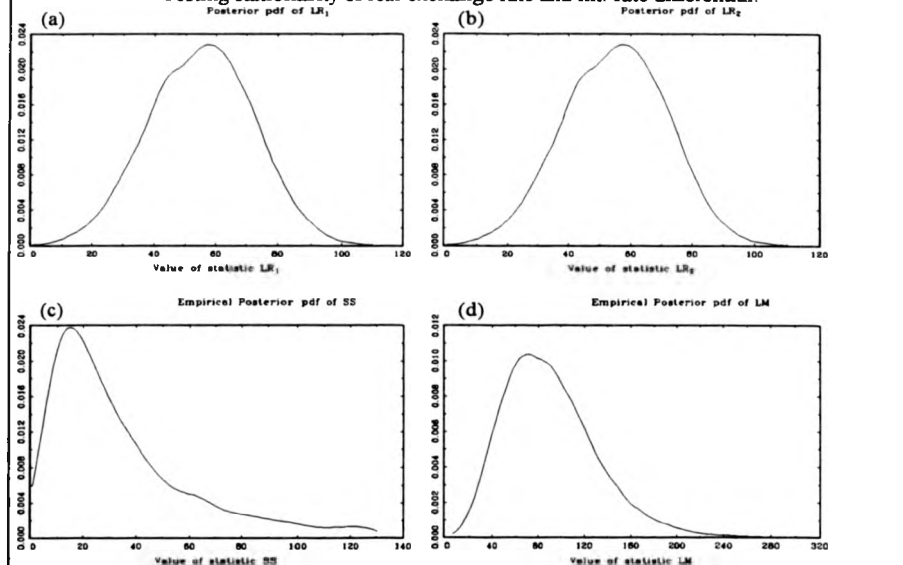


Figure (6.4.5). UK PPP/UIP. Co-int. rank=2.  $\omega_\phi=0.0$ .  
Testing stationarity of real exchange rate.

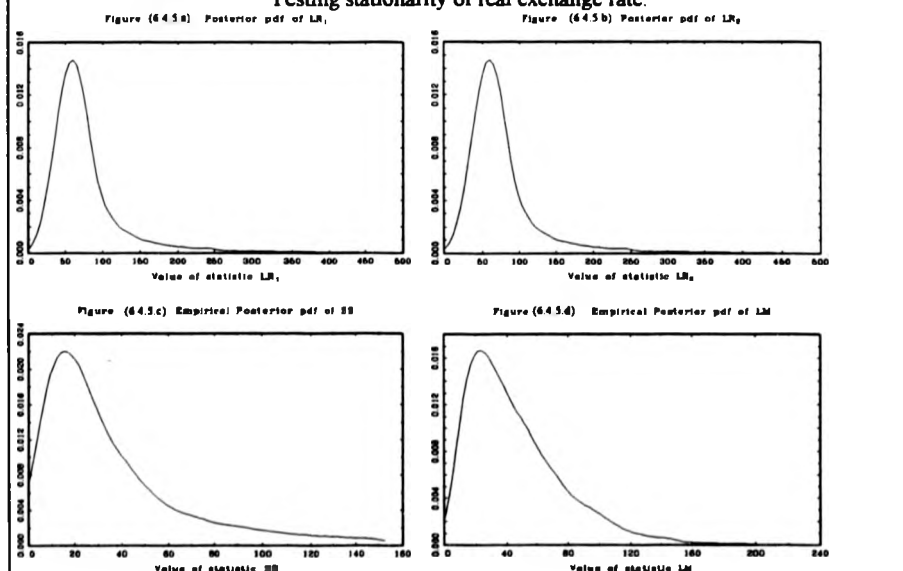


Figure (6.4.6). UK PPP/UIP. Coint. rank=2.  $\omega_{\phi}=0.0$ .  
Testing stationarity of int. rates differential

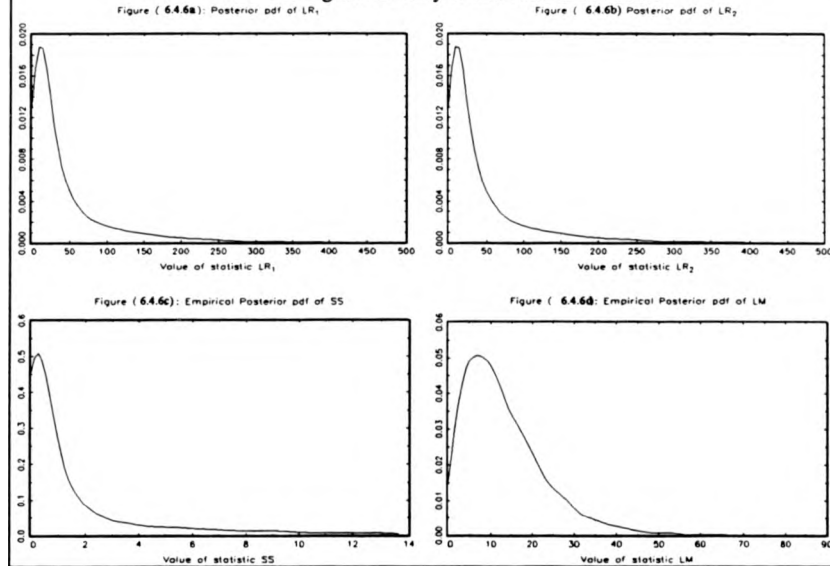


Figure (6.4.7). UK PPP/UIP. Coint. rank=2.  $\omega_{\phi}=0.1$ .  
95% HPD for  $\lambda_2$ : (0.02, 0.29)

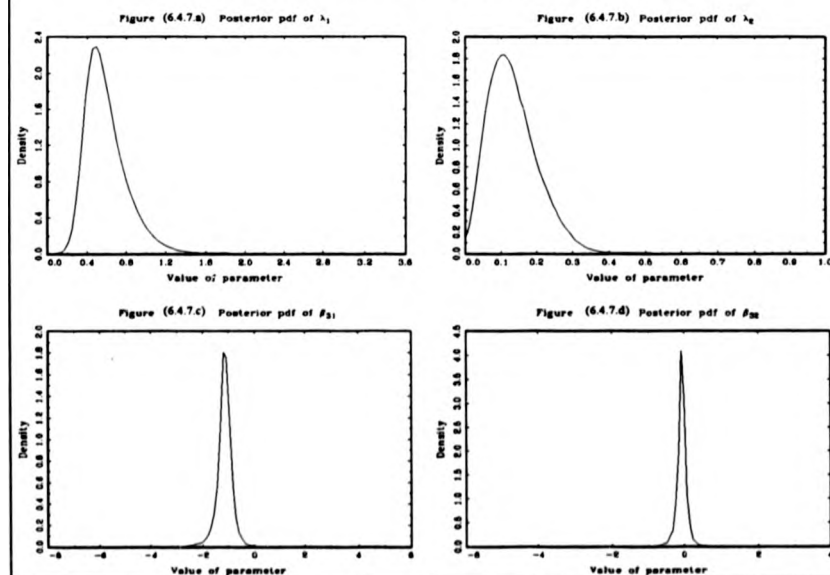


Figure (6.4.8) UK PPP/UIP. Coint.rank = 2.  $\omega_\phi = 0.1$

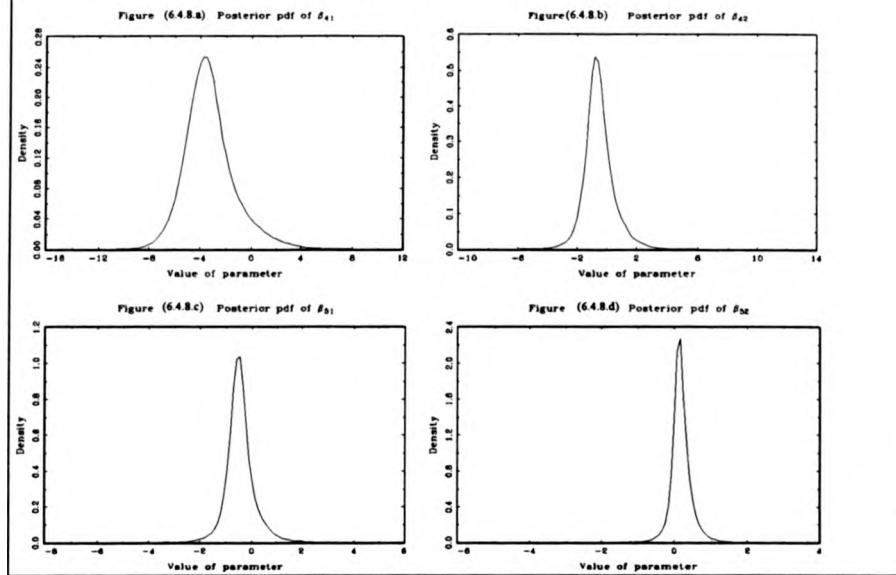


Figure (6.4.9) UK PPP/UIP. Coint.rank = 2.  $\omega_\phi = 0.1$ .  
Testing for stationarity of real exchange rate and int.rate differential

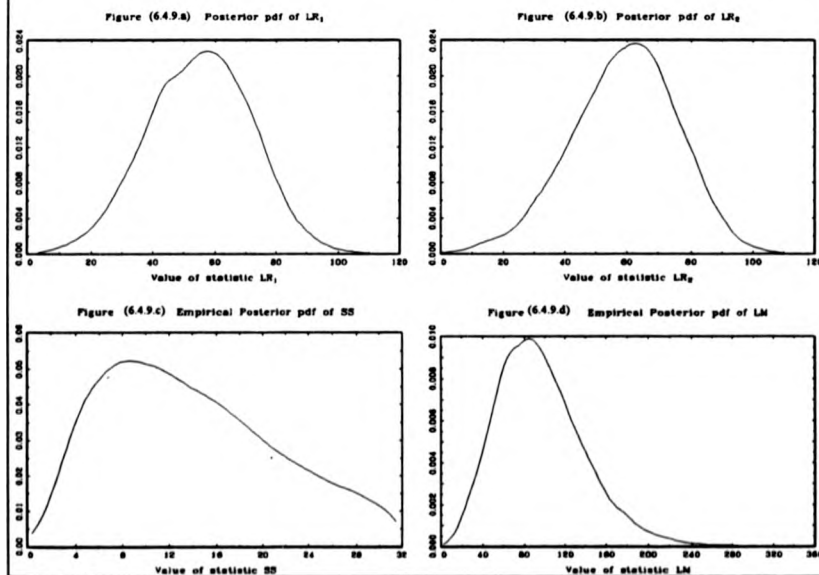


Figure (6.4.10) UK PPP/UIP. Coint.rank = 2.  $\omega_\phi = 0.1$   
Testing for stationarity of real exchange rate

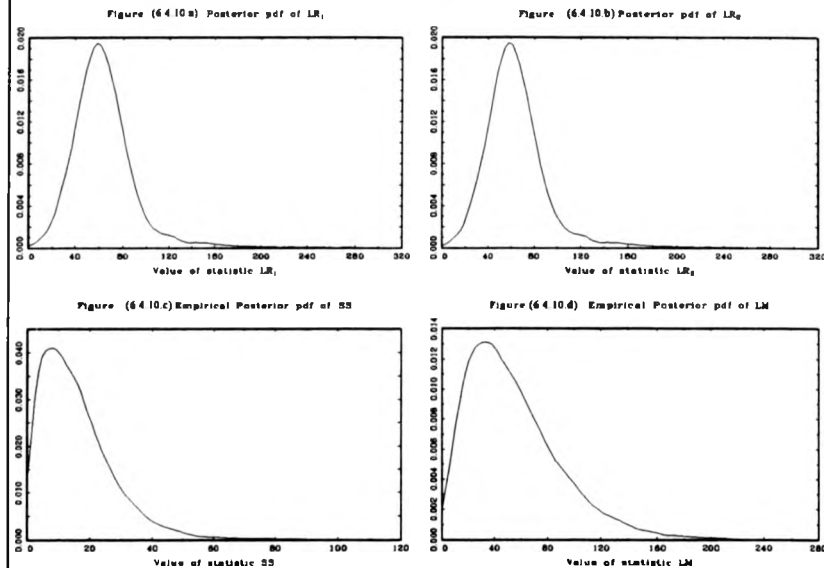
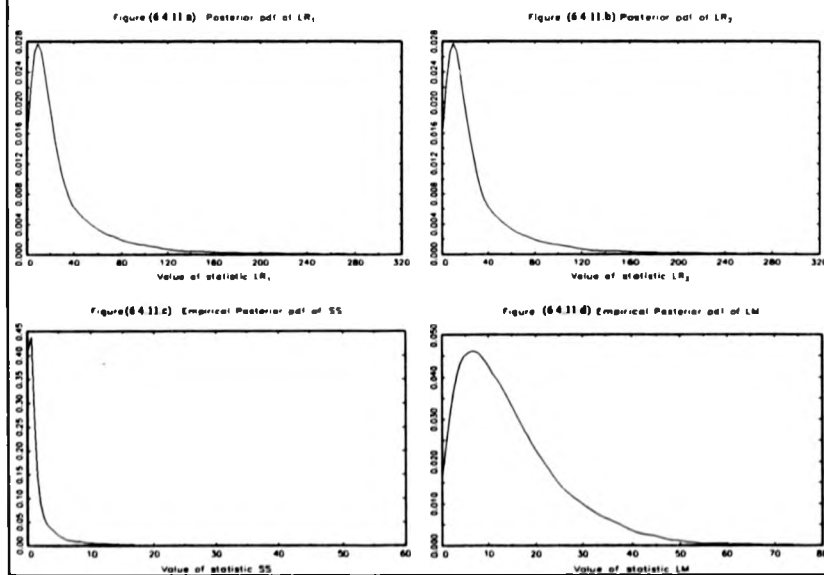


Figure (6.4.11) UK PPP/UIP. Coint.rank = 2.  $\omega_\phi = 0.1$   
Testing for stationarity of interest rate differential





## **Chapter 7: Concluding remarks**

In this thesis I have argued in favour of the use of Bayesian inferential techniques in the analysis of univariate and multivariate non-stationary linear time series models. Dealing with non-stationary series is somehow problematic, given that non-standard inferential results are involved. Such results hold only asymptotically and very little is known about finite sample properties. One of the main motivations behind the use of a Bayesian approach is then the possibility to avoid reliance on asymptotic distributional results.

Moreover, using Bayesian techniques, it is possible to incorporate prior beliefs in a clear way. A Bayesian study begins with a statement concerning the prior beliefs of the researcher. These beliefs are rigorously represented as a prior pdf, and they are mathematically revised, in the light of data evidence, to produce posterior distributions. On the other hand, in too many applications of the classical inferential techniques, the researcher tends to manipulate some results openly conflicting with prior views. An example of this practice is the widespread usage of larger sizes in the cointegration rank test to reconcile the test findings with the researcher's unstated priors in this regard. Under this point of view, the Bayesian approach is more 'honest', given that the prior beliefs are explicitly stated at the outset, and rigorously combined with the evidence coming from the data.

We have seen that the use of Bayesian techniques generates computational complications. Recent improvements in the use of Monte Carlo integration have widened the class of models which can be treated by means of Bayesian inferential techniques. The development of Markov chain methods in order to draw samples from the posterior distribution is clearly the most important advancement in this respect.

In this thesis I have conducted two different case studies with the aim to provide evidence of the applicability of Bayesian inferential techniques to univariate and multivariate unit root models.

The first of these cases, described in Chapter 4, is devoted to the analysis of univariate time series models, and develops a procedure to test whether a given quarterly macroeconomic time series presents seasonal and/or zero frequency unit roots. The inferential technique used is the evaluation of posterior odds ratios in order to compare hypotheses. The procedure is applied to a set of UK macroeconomic time series and the results are interesting from two different viewpoints. First of all, the results seem to be robust with respect to perturbations of the prior distributions; secondly, they conflict with some of the conclusions obtained on the basis of the classical asymptotic unit root tests.

The second case study, described in Chapter 6, develops a Bayesian procedure to conduct inference in potentially cointegrated *VAR* systems. Inference regards the number of cointegrating relationships being present in the data and their structural interpretation by testing and imposing over-identifying constraints on the cointegrating vector coefficients. The inferential procedure is based on the evaluation of highest posterior density confidence intervals, and on their use for decision making.

The procedure is applied to three different *VAR* systems providing interesting results: in the first of the three applications, regarding the system with money, income, bond and deposit interest rates for Denmark studied by Johansen and Juselius (1990), I find rank equal to one unequivocally, and I find unit income elasticity. On the other hand, I reject the second constraint imposed in the analysis of Johansen and Juselius (1990), i.e. the interest rates spread appearing in the long-run relationship.

The second application deals with the system of money, income, interest rate and inflation for Finland studied by Johansen and Juselius (1990); the cointegrating rank is found equal to three, and the structural interpretation of these vectors is achieved by testing and accepting the stationarity of inflation, of the interest rate, and of the log money-income ratio.

The third application regards the *UK* exchange rate data of Johansen and Juselius (1992): the cointegrating rank is unequivocally found equal to two and the constraints implying stationarity of the real exchange rate are squarely rejected, while those implying stationarity of the domestic-foreign interest rates differential are accepted. This seems to be the consequence of the *PPP* hypothesis failing to hold in the sample period being analysed. All these results have been proved to be robust with respect to different specifications for the prior distributions.

Some indications can be drawn from the case studies presented in this thesis. On a methodological point of view, some more work is needed in order to monitor the performances of the Markov chain Monte Carlo methods being implemented.

A crucial point is related to the problem of assessing whether convergence of the Markov chain sampling scheme to the target distribution has occurred or not. This problem has been dealt with in this thesis by comparing the sample means computed on the basis of early draws with the corresponding statistics obtained on the basis of the late draws of the sampling scheme. This approach is not completely satisfactory because it cannot detect the problems induced by the presence of multimodality of the target distribution. In such cases, the support of the distribution tends to become disconnected as the sample size increases, in this way precluding the convergence of the sampling scheme. Resorting to multiple chain samples with different starting values could be a sensible solution (Gelman and Rubin, 1992).

Another important problem is related to the evaluation of posterior odds ratios. In the recent literature (Newton and Raftery, 1994, Gelfand and Dey, 1994, Carlin and Chib, 1995) different ways have been proposed in order to evaluate posterior odds ratios, and their relative merits have been recently discussed by Kass and Raftery (1995). As documented in Section 4.5, in the applications presented in this thesis the hypotheses being compared have a nested structure, and therefore the Bayes factor computations seem not to involve any of the numerical complications described in Kass and Raftery (1995). More work is needed to assess the accuracy of different ways to compute Bayes factors.

As for the specific case studies presented in this thesis, some considerations seem necessary. The study of seasonal time series conducted in Chapter 4 considered only the alternatives of deterministic seasonality and the presence of seasonal unit roots. The analysis could be refined by taking into the consideration also the periodically varying coefficient models (see Franses, 1994) as a further viable alternative to modelling series with a seasonal pattern. A Bayesian procedure to discriminate between periodically varying coefficients and the presence of seasonal unit roots would render it possible to avoid treating the observations on each quarter of the year as a separate series, as in Franses (1994), and to rely on the estimation of models with very few degrees of freedom.

The cointegration analysis conducted in Chapter 6 can be extended in several directions. First of all, the impact of shrink to mean priors for the short run dynamics parameters is still to be assessed, together with the possibility to deal with systems of several variables.

Secondly, it would be interesting to obtain estimates of the univariate posterior distributions of the individual impulse response coefficients, simply by mapping the draws on the *ECM* parameters onto draws on the vector moving average representation. Also, the posterior distributions of the persistence profiles of

shocks affecting the structuralised cointegrating relationships can be obtained by simulation. Moreover, rational expectation hypotheses imposing non-linear constraints on the *VAR* parameters, such as the *UIP* model, can be tested on the basis of the finite sample posterior distributions of the statistics of interest.

Another interesting by-product of the cointegrating *VAR* analysis is given by the fact that the sample of the simulated disturbance variance-covariance matrices can be used to test the plausibility of the over-identifying non-linear constraints necessary to give the disturbance vector a "structural" interpretation, as in the *SVAR* analysis (see Bernanke, 1986, Sims, 1986, Giannini, 1992).

Thirdly, an interesting extension would be to allow for non-normality of the error terms, in order to allow the joint treatment of financial time series vectors, given that these data usually present fat tails. Alternative approaches are feasible in this respect. The easiest route could be to impose multivariate Student-*t* disturbances, and to work with a hierarchical model. In addition, or as an alternative to this strategy, a multivariate *GARCH* structure could be imposed on the disturbances.

## References

- Akaike, H. (1979): A Bayesian Extension of the Minimum AIC Procedure of Autoregressive Model Building, *Biometrika*, 66, 2, 237-242.
- Anderson, T. W. (1984): *An Introduction to Multivariate Statistical Analysis*, Wiley, New York.
- Andrews, D.W.K. (1991): Heteroskedasticity and autocorrelation consistent covariance matrix estimation, *Econometrica*, 59, 817-858.
- Andrews, D.W.K., and J.C. Monahan (1992): An improved heteroskedasticity and autocorrelation consistent covariance matrix estimation, *Econometrica*, 60, 953-966.
- Banerjee, D.F. Hendry, and Smith, G.W. (1986): Exploring equilibrium relationships in econometrics through static models: some Monte Carlo Evidence, *Oxford Bulletin of Economics and Statistics*, 52, 95-104.
- Banerjee, A., J. Dolado, J.W. Galbraith and D.F. Hendry (1993): *Co-Integration, Error Correction, and the Econometric Analysis of Non-Stationary Data*, Oxford University Press, Oxford.
- Barghava, A. (1986): On the Theory of Testing for Unit Roots in Observed Time Series, *Review of Economic Studies*, 52, 369-384.
- Bauwens, L., and M. Lubrano (1994): Identification Restrictions and Posterior Densities in Cointegrated Gaussian VAR Systems, CORE discussion paper 94-16, Universite' Catholique de Louvain.
- Baxter, M. and R.G. King (1993): Fiscal Policy in General Equilibrium, *American Economic Review*, 85, 315-334.
- Beaulieu, J.J., and J.A. Miron (1993): Seasonal Unit Roots in Aggregate U.S. Data, *Journal of Econometrics*, 55, 305-348.
- Berger, R.L. and D.F. Sinclair (1984): Testing Hypotheses Concerning Unions of Linear Subspaces, *Journal of the American Statistical Association*, 79, 158-163.
- Bernanke, B. (1986): Alternative Explanations of the Money-Income Correlation, *Carnegie-Rochester Conference Series on Public Policy*, 25, 49-100.
- Beveridge, S., and C.R. Nelson (1981): A New Approach to Decomposition of Economic Time Series into Permanent and Transitory Components, with Particular Attention to Measurement of the 'Business Cycle', *Journal of Monetary Economics*, 7, 151-174.
- Billingsley, P. (1968): *Convergence of Probability Measures*, Wiley, New York.

- Blanchard, O.J., and S. Fischer (1989): *Lectures in Macroeconomics*, Cambridge MA, MIT Press.
- Broemeling, L. (1985): *Bayesian Analysis of Linear Models*, New York, Marcel Dekker.
- Campbell, J. Y. (1994): Inspecting the Mechanism. An Analytical Approach to the Stochastic Growth Model, *Journal of Monetary Economics*, 33, 463-506.
- Campbell, J.Y., and P. Perron (1991): Pitfalls and Opportunities: What Economists Should Know about Unit Roots, *NBER Macroeconomic Annual*.
- Campbell, J.Y. and N.G. Mankiw (1987): Are Output Fluctuations Transitory?, *Quarterly Journal Of Economics*, 102, 857-880.
- Campbell J.Y., and R.J. Shiller (1987): Cointegration and Tests of Present Value Models, *Journal of Political Economy*, 95, 1062-1088.
- Cappuccio, N. and D. Lubian (1995): A Comparison of Alternative Approaches to Estimation and Inference in Structural Long Run Economic Equilibria, mimeo, Università di Padova, Italy.
- Carlin, B.P. and S. Chib (1995): Bayesian Model Choice via Markov Chain Monte Carlo Methods, *Journal of the Royal Statistical Society, B Series*, 57, 3, 473-484.
- Chan, K.S. (1993): Asymptotic Behavior of the Gibbs Sampler, *Journal of the American Statistical Association*, 88, 320-326.
- Chib, S. (1993): Bayes Regression with Autoregressive Errors, *Journal of Econometrics*, 58, 275-294.
- Chib, S. and E. Greenberg (1994a): Bayesian Inference in Regression Models with  $ARMA(p,q)$  Errors, *Journal of Econometrics*, 64, 183-206.
- Chib, S. and E. Greenberg (1994b): Markov Chain Monte Carlo Simulation Methods in Econometrics, mimeo, Washington University, St. Louis, MO. USA.
- Correia, I., J.C. Neves and S. Rebelo (1995): Business Cycle in a Small Open Economy, *European Economic Review*, forthcoming.
- Davidson, R. and J.G. Mac Kinnon (1994): *Estimation and Inference in Econometrics*, Oxford University Press, Oxford.
- DeJong, D.N., J.C. Nankervis, N.E. Savin and C.H. Whiteman (1992): Integration vs. Stationarity in Time Series, *Econometrica*, 60, 423-433.
- DeJong, D.N., and C.H. Whiteman (1989): Trends and Cycles as Unobserved Components in U.S. Real GNP: a Bayesian Perspective, *Proceeding of the*

*Business and Economic Statistics Section of the American Statistical Association.*

- DeJong, D.N., and C.H. Whiteman (1991): Trends and Random Walks in Macroeconomic Time Series. A Reconsideration Based on the Likelihood Principle, *Journal of Monetary Economics*, 28, 221-254.
- Devroye, L. (1986): *Non-Uniform Random Variate Generation*, New York, Springer-Verlag.
- Dhrymes, P.J. (1978): *Mathematics for Econometrics*, Springer Verlag, New York.
- Dickey, D.A. (1976): *Estimation and Testing of Non-Stationary Time Series*, Ph.D. Thesis, Iowa State University.
- Dickey, D.A. and W.A. Fuller (1979): Distributions of the Estimators for Autoregressive Time Series with a Unit Root, *Journal of the American Statistical Association*, 74, 427-431.
- Dickey, D.A. and W.A. Fuller (1981): Likelihood Ratio Statistics for Autoregressive Time Series with a Unit Root, *Econometrica*, 49, 1057-1072.
- Dickey, J.M. (1971): The weighted likelihood ratio, linear hypotheses on normal location parameters, *The Annals of Mathematical Statistics*, 42, 1, 204-223.
- Doan, T.R., R. Litterman and C.A. Sims (1984): Forecasting and Conditional Projection Using Realistic Prior Distributions, *Econometric Reviews*, 5, 1-56.
- Dreze, J.H. and J.F. Richard (1983): Bayesian Analysis of Simultaneous Equations Systems, in Z. Griliches, and M.D. Intriligator (eds.): *Handbook of Econometrics*, Vol. 1, North-Holland, Amsterdam.
- Durlauf, S.N. and P.C.B. Phillips (1988): Trends versus Random Walks in Time Series Analysis, *Econometrica*, 56, 1333-1354.
- Engle, R.F., C.W.J. Granger, S. Hylleberg and H.S. Lee (1993): Seasonal Cointegration: The Japanese Consumption Function, *Journal of Econometrics*, 55, 257-304.
- Engle, R. and C.W.J. Granger (1987): Co-integration and error correction: estimation, representation and testing, *Econometrica*, 55, 251-276.
- Franses, P.H. (1994): A Multivariate Approach to Modelling Univariate Seasonal Time Series, *Journal of Econometrics*, 63, 133-151.
- Fuller, W.A. (1976) : *Introduction to Statistical Time Series*, Wiley, New York.



- Gelfand, A.E. and A.F.M. Smith (1990): Sampling Based Approaches to Calculating Marginal Densities, *Journal of the American Statistical Association*, 85, 398-409.
- Gelfand, A.E., and D.K. Dey (1994): Bayesian Model Choice: Asymptotics and Small Sample Calculations, *Journal of the Royal Statistical Association, B Series*, 56, 3, 501-514.
- Gelman A. and D.B. Rubin (1992): Inference from Iterative Simulation Using Multiple Sequences, *Statistical Science*, 4, 457-472.
- Geman, S. and D. Geman (1984): Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-741.
- Geweke (1988): Antithetic Acceleration of Monte Carlo Integration in Bayesian Inference, *Journal of Econometrics*, 38, 73-89.
- Geweke, J., (1989): Bayesian Inference in Econometric Models Using Monte Carlo Integration, *Econometrica*, 57, 6, 1317-1339.
- Geweke, J., (1992): Evaluating the Accuracy of Sampling Based Approaches to the Calculation of Posterior Moments, in Berger, J.O., J.M. Bernardo, A.P. Dawid and A.F.M. Smith (eds.) *Bayesian Statistics*, Vol 4, Oxford University Press, Oxford.
- Geweke (1993): Bayesian Analysis of Reduced Rank Models, mimeo, University of Minnesota.
- Geweke, J., (1994): Priors for Macroeconomic Time Series and Their Application, *Econometric Theory*, 10, 609-632.
- Giannini, C. (1992): *Topics in Structural VAR Econometrics*, Springer-Verlag, Berlin.
- Granger C.W.J. (1981): Some Properties of Time Series Data and Their Use in Econometric Model Specification, *Journal of Econometrics*, 16, 101-30.
- Granger, C.W.J., and M.J. Morris (1976): Time Series Modelling and Interpretation, *Journal of the Royal Society*, 139, part 2, 246-257.
- Granger C.W.J., and P. Newbold (1974): *Spurious Regressions in Econometrics*, *Journal of Econometrics*, 2, 111-20.
- Hall, A. (1989): Testing for a Unit Root in the Presence of Moving Average Errors, *Biometrika*, 79, 49-56.
- Hall, P. and C.C. Heyde (1980): *Martingale Limit Theory and its Applications*, New York, Academic Press.

- Hamilton, J.D. (1994): *Time Series Analysis*, Princeton University Press, Princeton, New Jersey.
- Hammersley, J.M., and D.C. Handscomb (1964): *Monte Carlo Methods*, 1<sup>st</sup> Edition, London, Methuen.
- Hansen, L.P. (1982): Large Sample Properties of Generalized Method of Moments Estimation, *Econometrica*, 50, 1029-1054.
- Harvey, A.C., (1990): *Forecasting, Structural Time Series Models and The Kalman Filter*, Cambridge, Cambridge University Press.
- Hylleberg, S., R.F. Engle, C.W.J. Granger and B.S. Yoo (1990): Seasonal Integration and Cointegration, *Journal of Econometrics*, 44, 215-238.
- Jeffreys, H. (1946): An Invariant Form of the Prior Probability in Estimation Problems, *Proceeding of the Royal Society of London, Series A*, 186, 453-461.
- Jeffreys, H. (1961): *Theory of Probability*, 3rd Edition, London, Oxford University Press.
- Johansen, S. (1988): Statistical analysis of cointegrating vectors, *Journal of Economic Dynamics and Control*, 12, 231-254.
- Johansen, S. (1991): Estimation and Hypothesis Testing of Cointegrating Vectors in Gaussian Vector Autoregressive Models, *Econometrica*, 59, 1551-1580.
- Johansen, S. (1992): Determination of the Cointegration Rank in the Presence of a Linear Trend, *Oxford Bulletin of Economics and Statistics*, 54, 383-97.
- Johansen, S. (1995a): Identifying restrictions of linear equations, *Journal of Econometrics*, forthcoming.
- Johansen, S. (1995b): *Likelihood Based Inference on Cointegration in the Vector Autoregressive Model*, Oxford University Press, Oxford.
- Johansen, S. and K. Juselius (1990): Maximum Likelihood Estimation and Inference on Cointegration- with applications to the demand for money, *Oxford Bulletin of Economics and Statistics*, 52, 2, 169-210.
- Johansen, S. and K. Juselius (1992): Testing Structural Hypotheses in a Multivariate Cointegration Analysis of the PPP and the UIP for UK, *Journal of Econometrics*, 53, 211-244.
- Johansen, S. and K. Juselius (1992): Testing Structural Hypotheses in a Multivariate Cointegration Analysis of the PPP and the UIP for UK, *Journal of Econometrics*, 53, 211-244.

- Johansen, S. and K. Juselius (1994): Identification of the Long-Run and Short-Run Structure. An Application to the ISLM Model, *Journal of Econometrics*, 63, 7-36.
- Kass, R.E. and A.E. Raftery (1995): Bayes Factors, *Journal of the American Statistical Association*, 90, 773-95.
- King, R., C. Plosser and S. Rebelo: (1988): Production, Growth and Business Cycle I: The Basic Neoclassical Model, *Journal of Monetary Economics*, 21, 195-232.
- Kleibergen, F. and H. K. van Dijk (1994): On the Shape of the Likelihood/Posterior in Cointegrated Models, *Econometric Theory*, 10, 514-22.
- Kloek, T. and H.K. van Dijk (1978): Bayesian Estimates of Equation System Parameters: an Application of Integration by Monte Carlo, *Econometrica*, 46, 881-896.
- Koop, G. (1994): Recent Progress in Applied Bayesian Econometrics, *Journal of Economic Surveys*, 8, 1-34.
- Kydland, F.E., and E.C. Prescott (1982): Time to Build and Aggregate Fluctuations, *Econometrica*, 50, 1345-1370.
- Leamer, E.E. (1978): *Specification Searches*, Wiley, New York.
- Lindley, D.V. (1957): A Statistical Paradox, *Biometrika*, 44, 187-192.
- Lindley, D.V. (1961): The Use of Prior Probability Distributions in Statistical Inference and Decision, *Proceeding of the 4<sup>th</sup> Berkeley Symposium in Mathematical Statistics and Probability*, 1, 453-468.
- Lutkepohl, H. (1991): *Introduction to multiple time series analysis*, Berlin, Springer-Verlag.
- Mankiw, N.G. (1989): Real Business Cycles: a New Keynesian Perspective, *Journal of Economic Perspectives*, 3, 79-90.
- Mann, H.B. and A. Wald (1943): *On Stochastic Limit and Order Relationships*, *Annals of Mathematical Statistics*, 14, 217-77.
- Mengersen, K.L. and R.L. Tweedie (1993): Rates of Convergence of the Hastings and Metropolis Algorithms, mimeo, Colorado University.
- Mendoza, E. (1991): Real Business Cycle in a Small Open Economy, *American Economic Review*, 81, 797-818.
- Nelson, C.R. and C.I. Plosser (1982): Trend and Random Walks in Macroeconomic Time Series. Some Evidence and Implications, *Journal of Monetary Economics*, 10, 139-162.

- Nerlove, M., and M. Pinto (1984): Time and Frequency Domain Estimation of Time Series Models, manuscript, Dept. of Economics, University of Pennsylvania, PA.
- Newey, W.K., and K.D. West (1987): A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimator, *Econometrica*, 55, 703-708.
- Newton, M.A. and A.E. Raftery (1994): Approximate Bayesian Inference with the Weighted Likelihood Bootstrap, *Journal of the Royal Statistical Society, B Series*, 56, 1, 3-48.
- Osborn, D.R. (1990): A Survey of Seasonality in U.K. Macroeconomic Variables, *International Journal of Forecasting*, 6, 327-336.
- Ouliaris, S., J.Y. Park and P.C.B. Phillips (1988): Testing for a Unit Root in the Presence of a Maintained Trend, in Baldev Raj (Ed.): *Advances in Econometrics and Modelling*, Needham, MA, Kluwer Academic Publishers.
- Pantula, S.G. (1989): Testing for Unit Roots in Time Series Data, *Econometric Theory*, 5, 256-71.
- Park, J.Y., and P.C.B. Phillips (1988): Statistical Inference in Regressions with Integrated Processes, Part 1, *Econometric Theory*, 4, 468-497.
- Park, J.Y., and P.C.B. Phillips (1989): Statistical Inference in Regressions with Integrated Processes, Part 2, *Econometric Theory*, 5, 95-131.
- Perks, F.G.A. (1947): Some Observations on Inverse Probability Including a New Indifference Rule, *Journal of the Institute of Actuaries*, 73, 285-334.
- Perron, P. (1988): Trends and Random Walks In Macroeconomic Time Series: Further Evidence From A New Approach, *Journal Of Economic Dynamics and Control*, 12, 297-332.
- Perron, P. (1989): The Great Crash, The Oil Price Shock and The Unit Root Hypothesis, *Econometrica*, 57, 1361-1401.
- Perron, P. and P.C.B. Phillips (1987): Does Gnp Have A Unit Root? A Re-evaluation, *Economic Letters*, 23, 139-145.
- Pesaran, M.H. and Y. Shin (1995): Cointegration and the Speed of Convergence to Equilibrium, *Journal of Econometrics*, forthcoming.
- Phillips, P.C.B. (1983): Marginal Densities of Instrumental Variable Estimators in the General Single Equation Case, *Advances in Econometrics*, 2, 1-24.
- Phillips, P.C.B. (1986): Understanding Spurious Regressions in Econometrics, *Journal of Econometrics*, 33, 311-40.

- Phillips, P.C.B. (1987): Time Series Regression with a Unit Root, *Econometrica*, 55, 277-301.
- Phillips, P.C.B. (1991a): To Criticize the Critics: an Objective Bayesian Analysis of Stochastic Trends, *Journal of Applied Econometrics*, 6, 435-73.
- Phillips, P.C.B. (1991b): Optimal Inference in Co-Integrated Systems, *Econometrica*, 59, 282-306.
- Phillips, P.C.B. (1994): Some exact Distribution Theory for Maximum Likelihood Estimators of Cointegrating Coefficients in Error Correction Models, *Econometrica*, 62, 73-93.
- Phillips, P.C.B. and S.N. Durlauf (1986): Multiple Time Series Regression with Integrated Processes, *Review of Economic Studies*, 53, 473-95.
- Phillips, P.C.B., and B.E. Hansen (1990): Statistical Inference in Instrumental Variables Regression with I(1) Processes, *Review of Economic Studies*, 57, 99-125.
- Phillips, P.C.B. and P. Perron (1988): Testing for a Unit Root in Time Series Regression, *Biometrika*, 75,2, 335-346.
- Plosser, C.I. (1989): Understanding Real Business Cycles, *Journal of Economic Perspectives*, 3, 51-77.
- Prescott, C.E. (1986): Theory Ahead of Business Cycle Measurement, in *Carnegie-Rochester Series on Public Policy*, 125, 11-66.
- Raiffa, H.A. and R.S. Schlaifer (1961): *Applied Statistical Decision Theory*, Harvard University, Boston.
- Ripley, B.D., (1987): *Stochastic Simulation*, Wiley, New York.
- Ritter, C. and M.A. Tanner (1992): The Gibbs Stopper and the Griddy Gibbs Sampler, *Journal of the American Statistical Association*, 87, 861-8.
- Roberts, G.O. and A.F.M. Smith (1994): Simple Conditions for the Convergence of the Gibbs Sampler and Metropolis-Hastings Algorithms, *Stochastic Processes and Their Applications*, 49, 207-16.
- Rubinstein, R.Y. (1985): *Simulation and the Monte Carlo Method*, Wiley, New York.
- Said, S.A and D.A. Dickey (1984): Testing For Unit Roots In Autoregressive-Moving Average Models Of Unknown Order, *Biometrika*, 71, 599, 608.
- Sargan, J.D. (1988): *Lectures on advanced econometric theory*, Oxford, Basil Blackwell.

- Sargan, J.D. and A. Barghava (1983): Testing Residuals from Least Squares Regression for Being Generated by the Gaussian Random-Walk, *Econometrica*, 51, 153-74.
- Schmidt, P., and P.C.B. Phillips (1992): Testing for a Unit Root in the Presence of Deterministic Trends, *Oxford Bulletin of Economics and Statistics*, 54, 257-287.
- Schotman, P. and H.K. Van Dijk (1991 a): A Bayesian Analysis of the Unit Root in Real Exchange Rates, *Journal of Econometrics*, 49, 195-238.
- Schotman, P. and H.K. Van Dijk (1991 b): On Bayesian Routes to Unit Roots, *Journal of Applied Econometrics*, 6, 387-401.
- Schotman, P. and H.K. Van Dijk (1992): Posterior Analysis of Possibly Integrated Series with an Application to Real GNP, in P. Caines, J. Geweke and M. Taquq (eds.) *New Directions in Time Series Analysis*, Part II. Berlin: Springer-Verlag.
- Schwert, G.W. (1988): Effects of Model Misspecification on Tests for Unit Roots in Macroeconomic Data, *Journal of Monetary Economics*, 20, 73-103.
- Schwert, G.W. (1989): Testing for Unit Roots : A Monte Carlo Investigation, *Journal of Business and Economic Statistics*, 6, 13-28.
- Silverman, B.W. (1986): *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.
- Sims, C.A. (1974): Seasonality in Regression, *Journal of the American Statistical Association*, 69, 618-626.
- Sims, C.A. (1980): Macroeconomics and Reality, *Econometrica*, 48, 1-48.
- Sims, C.A. (1986): Are Forecasting Models Usable for Policy Analysis?, *Quarterly Review of the Federal Reserve Bank of Minneapolis*, winter, 2-16.
- Sims, C.A. (1988): Bayesian Skepticism on Unit Root Econometrics, *Journal of Economic Dynamics and Control*, 12, 463-474.
- Sims, C.A. (1989): Modeling Trends, discussion paper, Institute for Empirical Macroeconomics, and Federal Reserve Bank, Minneapolis.
- Sims, C.A. and H. Uhlig (1991): Understanding Unit Rooters: a Helicopter Tour, *Econometrica*, 59, 1591-1599.
- Sims, C.A., J. Stock and M. Watson (1990): Inference in Linear Time Series Models with Some Unit Roots, *Econometrica*, 58, 113-144.
- Smith, A.F.M., and G.O. Roberts (1993): Bayesian Computation Via the Gibbs Sampler and Related Markov-Chain Monte Carlo Methods, *Journal of The Royal Statistical Society*, B, 55,1, with discussions.

- Solo, V. (1984): The Order of Differencing in ARMA Models, *Journal of the American Statistical Association*, 79, 916-921.
- Spanos, A. (1986): *Statistical Foundations of Econometric Modelling*, Cambridge, Cambridge University Press.
- Stock, J.H. (1987): Asymptotic Properties of Least Squares Estimators of Co-Integrating Vectors, *Econometrica*, 55, 1035-56.
- Stock, J.H., and M. W. Watson (1988): Testing for Common Trends, *Journal of the American Statistical Association*, 83, 1097-1107.
- Theil, H. and S. Goldberger (1961): On Pure and Mixed Statistical Estimation in Economics, *International Economic Review*, 2, 65-78.
- Tierney, L. (1991): Exploring Posterior Distributions Using Markov Chains, University of Minnesota School of Statistics.
- Tierney, L. (1994): Markov Chains for Exploring Posterior Distributions (with discussion), *Annals of Statistics*, 21, 1701-1724.
- Tierney, L. and J.B. Kadane (1986): Accurate Approximations for Posterior Moments and Marginal Densities, *Journal of the American Statistic Association*, 79, 916-921.
- van Dijk, H.K. (1986): Some Advances in Bayesian Estimation Methods Using Monte Carlo Integration, *Advances in Econometrics*, 6, 215-261.
- Wallis, K.F. (1974): Seasonal Adjustment and Relations Between Variables, *Journal of the American Statistical Association*, 69, 18-32.
- West, K.D. (1988): Asymptotic Normality when Regressors Have a Unit Root, *Econometrica*, 56, 1397-1417.
- Yule, G.U. (1926): Why Do We Sometimes Get Nonsense Correlations Between Time Series? A Study in Sampling and the Nature of Time Series, *Journal of the Royal Statistical Society*, 89, 1-64.
- Zellner, A. (1971): *An Introduction to Bayesian Inference in Econometrics*, Wiley, New York.